An Image Analysis and Machine Learning Approach to Measuring the Quality of Individual Colonoscopy Procedures

A thesis submitted to the University of Dublin
in fulfilment of the requirements for the degree of
Doctor in Philosophy

by

Mirko Arnold

24th September 2012

Trinity College Dublin
Faculty of Engineering, Mathematics and Science
School of Computer Science and Statistics

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Mirko Arnold

25 September 2012

Abstract

The measurement of quality in colonoscopy is an active topic in medical research. Studies report significant miss rates in the detection of colorectal lesions. This has raised the concern among gastroenterologists that the present mechanisms for quality assurance are insufficient. The current clinical practice of quality assurance is based on long term statistics, while the quality of individual colonoscopy procedures is judged by self-assessment. For training and auditing, there exist validated subjective assessment methods, which involve the rating of procedures by trained experts using predefined assessment forms. We focus our research on one such assessment method, the Direct Observation of Procedure and Skill (DOPS), developed by the Joint Advisory Group on Gastrointestinal Endoscopy (JAG) in the UK. One of our main objectives is to investigate to what degree the JAG DOPS assessment can be automated.

We have developed a system to automatically measure the quality of colonoscopy procedures according to JAG DOPS criteria and have performed a pilot validation of the system using two trained clinical assessors. The system is based on two different types of data: video data from the endoscopic camera and measurements of the longitudinal and circular motion of the shaft of the endoscope outside the anus. We have developed a number of algorithms that measure different quality related characteristics in endoscopic images and complete colonoscopy procedures. While the development of these measures is oriented towards the overall objective of assessing JAG DOPS criteria, each measure represents clinically relevant image or procedure characteristics on its own.

For single images, we propose methods for the measurement of the clarity of the endoscopic field of view, the position and presence of the lumen, the quality of luminal views and the distance to the nearest bend in the colon. The image measures are based on models which are trained using a universal machine learning framework involving automatic feature selection and different variants of support vector machines. The quality of the features is enhanced by a number of pre-processing steps, most notably by a novel algorithm for detection and inpainting of specular highlights.

We estimate the depth of insertion of the endoscope using the measurements of the motion sensor, which allows us to divide the colon into a number of spatial segments. This representation is the basis for the development of novel measures of colonoscopy procedure characteristics, reflecting handling patterns and summarising image based measures over the course of complete procedures. The individual procedure measures are then used as features for the training of predictive models for the automatic assessment of JAG DOPS criteria.

Acknowledgements

First of all, I would like to thank my beloved wife Ola and my family for their patience and support during the last years, especially during the sometimes difficult periods and little crises.

I am also grateful to my supervisor Gerard Lacey for his guidance and many helpful discussions.

Very important was also the support of my colleagues in Trinity College, helping me to understanding the technical and academic challenges or, sometimes, carrying all that heavy equipment around. A special mention deserve Anarta Ghosh, Stefan Ameling, Darren Caulfield, Fernando Vilariño, Peter Redmond and Jorge Bernal. Thank you very much for everything!

The experiments in this thesis would not have been possible without the support of Hugh Mulcahy, Stephen Patchett, Glen Doherty and Chris Steele. My sincere thanks also go out to all participating endoscopists.

This work was carried out as part of the Endoview project with the funding support of Enterprise Ireland.

Mirko Arnold
September 2012

Contents

List of Figures

Chapter 1

Introduction

Colonoscopy is considered the gold standard for colorectal cancer screening. In addition to a thorough visualisation of the large intestine, it is possible to directly take tissue samples or remove colorectal polyps. Detection and removal of such polyps can prevent them from developing into cancer, which is why many organisations recommend regular colorectal screening for people over a certain age [1, 2], e.g., the American college of gastroenterology (ACG) recommended colonoscopy screening every 10 years from the age of 50 in a recent guideline [3]. With the commencement of screening programs for asymptomatic patients in more and more countries [4], the number of colonoscopy procedures is constantly growing.

While colonoscopy can reduce the incidence of colorectal cancer, it does not eliminate the risk completely. In fact, studies have shown that in practice, the percentage of patients developing colorectal cancer shortly after having undergone colonoscopy screening is between 2 % and 6 % [5]. Since the development of small polyps into cancer is known to be a gradual process that evolves slowly over years [6], this means that a significant number of polyps remain undetected despite colonoscopy screening.

The cause for these miss rates is not yet well understood. It is likely to be a combination of a number of factors. Screening technique and bowel preparation play an important role, as they influence the amount of colonic mucosa that is visualised. Perceptual and cognitive aspects have to be taken into account where visualised polyps are not correctly identified by the performing endoscopist. All these factors contribute to the overall quality of a colonoscopy procedure, in the sense that the quality of a procedure is optimal if all colorectal lesions were detected and treated appropriately.

In practice, the quality of colonoscopy procedures is assessed subjectively by the performing endoscopist. Additionally, a number of measures are recorded for statistical evaluation. Examples of such measures are the average withdrawal time or the adenoma de-

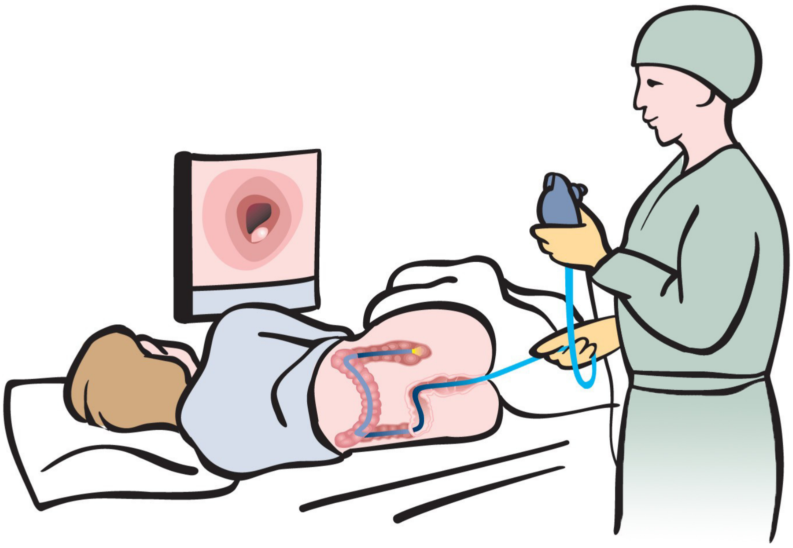


Figure 1.1: Depiction of a colonoscopy procedure.

tection rate (the ratio of patients undergoing colonoscopy in whom adenomatous polyps[1] were found [10]). These measures, however, can only assess the average performance of the endoscopists and long-term improvements of their skills.

For individual procedures there exist subjective quality measures such as the direct observation and assessment of the procedures by one or more experts. A number of validated assessment forms are used in such scenarios [11, 12]. Due to the cost and limited availability of trained assessors it is impractical to use this form of quality assessment more than for occasional audits or as part of the examination of trainees. The routine assessment of individual procedures is currently the task of the performing endoscopist alone.

Due to the drawbacks of self-assessment and the limited capabilities of long-term measures, the issue of quality assessment is regularly discussed in the gastroenterology community. Section 2.2 deals with this subject and contains a detailed discussion on quality measures for colonoscopy.

1.1 Objective

The open issues in quality assessment for endoscopic procedures motivate our research into whether the assessment of endoscopic skill and procedure quality can be automated. Our objective can be summarised with the following research question:

[1] Adenomatous polyps are benign tumours, which have the potential to develop into cancer. For details on types of polyps and colorectal lesions see, e.g., [7, 8, 9]

- To what degree can the criteria of current subjective quality assessment systems for individual colonoscopy procedures, such as JAG DOPS, be measured automatically from endoscopic video information and measurements of endoscope motion outside the anus?

JAG DOPS stands for the subjective assessment protocol DOPS (direct observation of procedure and skill) developed by the Joint Advisory Group on Gastrointestinal Endoscopy (JAG) in the UK. We consider this a particularly relevant example of an assessment tool for individual colonoscopy procedures. The motion measurements are obtained from a prototype motion sensor developed by our research group, which we will introduce in Chapter 3. The sensor measures longitudinal and circular motion of the shaft of the endoscope outside the anus.

For our particular approach, the research question can be subdivided into the following questions:

- Which characteristics of colonoscopy procedures are relevant for the assessment of JAG DOPS criteria?
- Can characteristics of endoscopic images and endoscope motion be combined for the measurement of these procedure characteristics?
- Can the procedure characteristics be mapped to JAG DOPS assessment criteria?
- How does the so obtained automatic assessment compare to the manual assessment by trained experts?

The questions indicate that we particularly focus our research on the assessment of *individual* procedures as opposed to long term assessment of endoscopists. Nevertheless, we also discuss the potential of the developed measures as long term indicators of the skill level of endoscopists.

It is important to note that we address the *measurement* of the procedure quality as opposed to quality *enhancement* (e.g., automated detection or classification of polyps) or the assessment of diagnostic abilities of the endoscopist. We focus on measures that allow to assess whether the prerequisites are sufficient for high quality bowel cancer screening.

It is envisaged that the outcome of this research will provide a basis for the development of objective quality measures, which are measured routinely and automatically, in order to ensure a high level of quality in gastroenterology services.

1.2 Approach

As our focus lies on the assessment of *individual* colonoscopy procedures, we commence our research by investigating existing methods of assessment for individual procedures. Current methods use direct observation by experts who rate the performance of the endoscopist according to validated assessment forms. After discussing examples of such assessment methods, we choose JAG DOPS as a particularly relevant form of assessment. We analyse the JAG DOPS assessment criteria and identify, which of them can potentially be measured automatically.

For this automated measurement, there are two different types of data we consider in this thesis. One is video data from the endoscopic camera. The other is data obtained from a motion sensor device. The sensor allows measurement of the longitudinal and circular motion of the shaft of the endoscope outside the anus. We investigate on what basis the chosen DOPS criteria are assessed in practice and identify patterns of image features and endoscope motion to be considered for modelling these underlying characteristics. The development of these measures is oriented towards the objective of assessing DOPS criteria, while each of the measures is intended as an independent representation of relevant, quality related image or procedure characteristics.

The individual measures are organised into two levels. The first contains all measures characterising single images, while the second level describes characteristics of the complete procedures. The measures of single images are incorporated in the second level by summarising their behaviour over the course of the procedure.

For the procedure measures and their mapping to DOPS criteria we use data obtained from an experiment in which endoscopists performed screening procedures on a colonoscopy training model. The data contains videos from the endoscope camera together with motion sensor readings, information on the experience and performance of the endoscopists and ratings of the procedures by two trained experts according to JAG DOPS criteria. The motion sensor readings are combined with the image based characteristics to measure a number of endoscope handling patterns. Furthermore, we use the recorded longitudinal motion of the shaft of the endoscope to estimate the depth of insertion of the endoscope. All image based characteristics can therefore be analysed for their behaviour over time and depth of insertion.

This combination of image and endoscope motion characteristics results in a large set of measures, describing colonoscopy procedures in great detail. We use subsets of these measures as features for the training of regression models for each of the chosen JAG DOPS criteria. We evaluate the proposed method by comparing the model predictions to the ratings of the trained experts.

1.3 Contributions

In this thesis, we will describe a number of contributions to the state of the art for the measurement of characteristics of endoscopic images, which are especially relevant for quality assessment. These characteristics are:

- The clarity of the endoscopic field of view
- The position of the lumen
- The presence of the lumen
- The quality of the luminal view
- The degree to which luminal view quality is compromised by the vicinity to the next bend

Other novel algorithms we propose here for analysing and enhancing single endoscopic images are:

- Detection of specular highlights
- Inpainting of specular highlights

Both methods are used in most of the image measures to improve their accuracy.

Another contribution is the design and analysis of a large experiment involving the assessment of 28 endoscopists performing a colonoscopy procedure on a simulator. Simultaneous recording of endoscope video data, orifice motion sensor data, eye tracking data and video data from an external perspective, together with attributes of the endoscopists and rating of the procedures by trained experts, make this experiment both novel and technically challenging. The resulting data set is of great value and used within this thesis to evaluate the proposed quality measures.

Given the video and endoscope motion data from the experiment, we propose a number of automatic measures of characteristics of whole procedures. These contributions can be summarised under the following headings:

- An estimate of the depth of insertion of the endoscope
- Insertion and withdrawal time
- Characteristics of the speed of endoscope handling
- A measure for pushing without a clear endoscopic field of view
- Measurement of stationary time during insertion

- A measure of attempted loop resolution
- Measures summarising the image measures over the course of whole procedures

The mapping of the procedure characteristics to clinically established quality criteria of colonoscopy procedures forms another contribution. The joint use of video and motion sensor data is a unique approach and, with the mapping to JAG DOPS criteria, we are the first to compare to an established quality assessment tool for individual procedures.

1.4 Layout of the Thesis

The thesis is laid out as follows. In Chapter 2, we provide detailed background information on colonoscopy practice and the related quality criteria. We review the current standards in quality assessment for colonoscopy procedures and explain the chosen direction for our research. Furthermore, we present a detailed review of previous research on the automatic analysis of colonoscopy video. Chapter 3 lists and explains our available sources of data and analyses the potential of each JAG DOPS criterion to be measured automatically given this data. Chapter 4 contains detailed descriptions and evaluation of the proposed methods for measuring image-level characteristics of colonoscopy procedures. In Chapter 5 we report on the design and implementation of an experiment for the collection of video and motion data from colonoscopy procedures, paired with the DOPS assessment of the procedures by two domain experts. We assess the reliability of this data by analysing the agreement between the raters and analyse the degree of association between the DOPS ratings and other recorded characteristics of the endoscopists and the procedures. Subsequently, we introduce and discuss a number of novel measures of characteristics of complete colonoscopy procedures in Chapter 6. Finally, we describe and evaluate our approach to mapping these procedure measures to the DOPS criteria. We then summarise and discuss the findings of the research programme in Chapter 7 before we conclude the thesis with an outlook on possible future research directions.

1.5 List of Publications

In the course of this work, intermediate results were published in the following papers:

- Mirko Arnold, Stefan Ameling, Anarta Ghosh and Gerard Lacey, Quality Improvement of Endoscopy Videos, Proc. IASTED International Conference on Biomedical Engineering, 2011

- Mirko Arnold, Anarta Ghosh, Stefan Ameling and Gerard Lacey, Automatic Segmentation and Inpainting of Specular Highlights for Endoscopic Imaging, EURASIP Journal on Image and Video Processing, 2010
- Mirko Arnold, Anarta Ghosh, Stephen Patchett, Hugh Mulcahy, Gerard Lacey, Indistinct Frame Detection in Colonoscopy Videos, 13th IMVIP conference, Dublin, 2009, pp47 - 52

Chapter 2

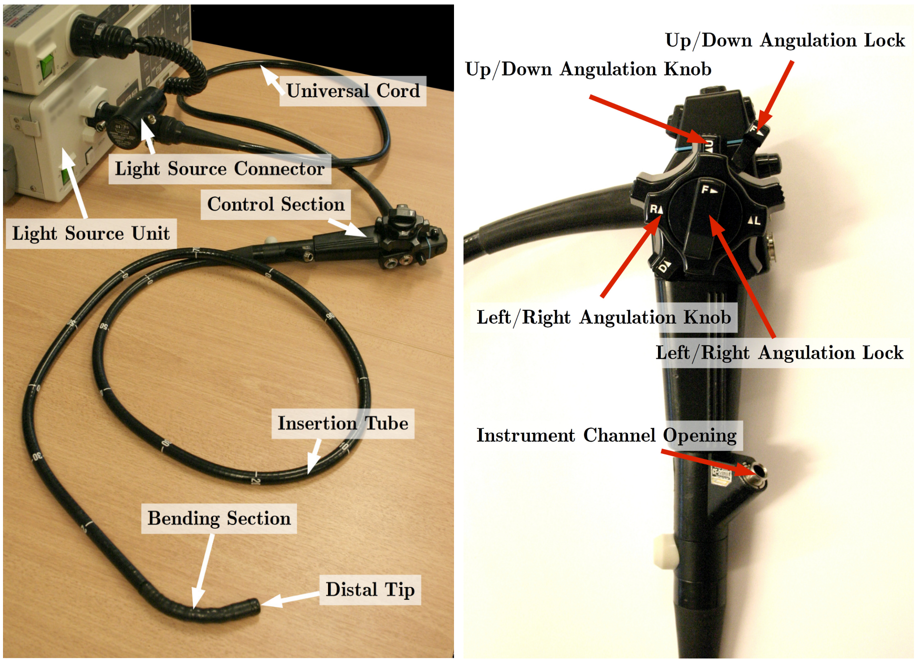
Background

2.1 Principles of Colonoscopy

Colorectal cancer is the 4th leading cause of cancer related deaths all over the world, behind lung, stomach and liver cancer [13]. It usually develops out of neoplastic polyps in the colon or rectum [14]. The cancer itself starts as localised colorectal cancer on the bowel wall. It can then spread into lymph nodes and develop distant metastases. If colorectal polyps are detected and removed, cancer can be prevented. When the cancer is diagnosed in the localised stage, the 5 year survival rate is still 90%, while it drops significantly at later stages. That is why many authorities recommend regular screening of people from a certain age, e.g., the American college of gastroenterology (ACG) recommended colonoscopy screening every 10 years from the age of 50 in a recent guideline [3].

Colonoscopy (or video colonoscopy) is a minimally invasive video screening of the large intestine [15, 7]. It is performed using an endoscope, a flexible tubular instrument with a camera and an illumination unit at the tip. Figure 2.1 shows a typical endoscope and a detailed view of its control section. The endoscope is inserted through the anus and passed through the rectum and the colon until the caecum is reached (see Fig. 2.2 for a schematic of the anatomy). The main screening is then performed during a second phase, during which the endoscope is gradually withdrawn. The signal from the video camera is shown on a screen to allow the endoscopist to thoroughly inspect the intestinal mucosa (the surface of the intestinal wall) for abnormalities such as polyps or other lesions. During a colonoscopy procedure the endoscopist can also collect tissue samples and perform therapeutic operations, e.g., remove polyps.

Other screening techniques exist, such as, for example, wireless capsule endoscopy (WCE) [16], or virtual colonoscopy (VC) [17], also referred to as computed tomography colonography (CTC). In WCE, the patient swallows a capsule containing a camera system. The capsule gets transported through the body by peristalsis. It transmits video data



(a) Components of an endoscope.

(b) Control section of an endoscope.

Figure 2.1: Illustration of the components of a typical endoscope.

wirelessly to a receiver that the patient carries. WCE is less invasive than conventional video colonoscopy and additionally allows screening of the small intestine. However, the capsule can currently not be controlled inside the body and the analysis of WCE video is difficult and very time consuming [18, 19, 20], due to the long duration of the obtained videos (on the order of hours). For latest generation colon capsules, however, a high sensitivity in finding polyps has been reported [21].

Virtual colonoscopy creates a 3D model of the colon from information obtained from computed tomography (CT) or magnetic resonance imaging (MRI) data. It has achieved promising results in the diagnosis of several diseases [22]. A comparison of the efficiency of virtual and video colonoscopy in detecting colonic polyps can be found in [23, 24, 25, 26]. A major drawback of virtual colonoscopy is its difficulty in detecting flat polyps, which are considered more dangerous than the more easily detectable peduncular types [9]. The level of acceptance for VC as a valuable technique for colorectal cancer screening varies among experts and institutions. The Centers for Medicare & Medicaid Services (CMS) in the U.S.A., for example, have recently decided that VC will no longer be covered by Medicare, stating that the current evidence supporting VC as an appropriate colorectal cancer screening test is “inadequate” [27]. Among the different alternatives for

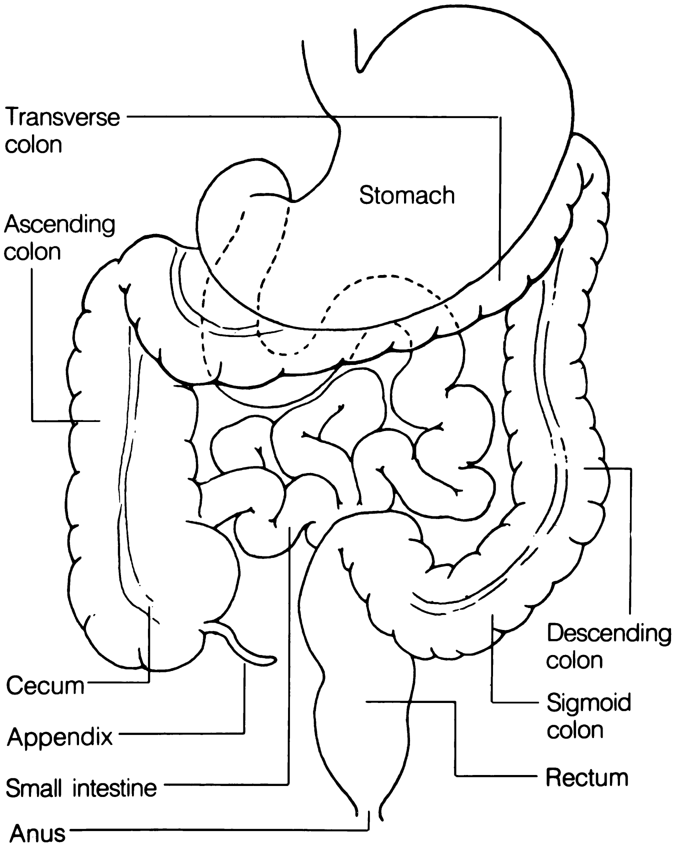


Figure 2.2: Illustration of the lower intestine.

colorectal cancer screening, video colonoscopy is the only method that allows therapeutic interventions, such as polypectomy (the removal of polyps) or the collection of tissue samples.

2.2 Quality in Colonoscopy

The importance of quality control in colonoscopy has been highlighted in recent publications [28, 29]. In an official recommendation of the U.S. Multi-Society Task Force on Colorectal Cancer, it is stated that the members “anticipate that the quality of colonoscopy will be among the most important issues surrounding its use” [30]. Also recently, David Lieberman, past president of the American Society for Gastrointestinal Endoscopy, stated that “it is time for all endoscopists to routinely measure quality indicators in their practice and strive for continuous quality improvement” [28].

Studies suggest that a significant number of lesions are missed during colonoscopy screening procedures [31, 32, 33]. In such a study it has been found that polyp miss rates can be as high as 36 % [34]. The reasons for these miss rates have not yet been sufficiently investigated. One possible reason may be poor bowel preparation [35, 36], which may result in the presence of large amounts of intestinal contents. These fluid or solid materials may cover areas of the intestinal mucosa where lesions are present. Another reason may be the lack of expertise of the performing physician [37]. The orientation and navigation inside the colon is not an easy task and requires extensive training and experience. It can therefore happen that some parts of the colon are not inspected properly.

As the anatomical complexity varies significantly from patient to patient, the expertise of endoscopists is best assessed in studies involving tandem colonoscopy (or back to back colonoscopy), i.e., two colonoscopy procedures performed in the same patient, usually on the same day. Such studies are conducted with the intention to finding correlations between certain procedural characteristics and performance measures. This can be, for example, the *polyp miss rate* (the number of polyps missed in one procedure compared to the overall found polyps [33, 38]).

Apart from tandem colonoscopies, studies are also often based on the adenoma detection rate. For example, it has been shown that the adenoma detection rate correlates significantly with the time spent during withdrawal, i.e., longer average withdrawal times yield higher detection rates [39, 40, 41]. These findings have led to the development of new quality guidelines for colonoscopy, e.g., [14], which are expected to yield a higher quality standard. Several quality indicators are currently used, such as the time spent for the withdrawal of the endoscope or whether a certain part of the colon has been reached.

Recently, the validity of some of the quality measures that were widely accepted, such as the withdrawal time, has been questioned [42]. What is still often used as a benchmark score in recent publications is the adenoma detection rate of endoscopists (e.g., in [43, 37]). However, this measure can also only be determined from a reasonably high number of procedures. Furthermore, it is dependent on the patient population. For example does the likelihood of developing adenoma increase with age. Therefore, an endoscopist whose patients on average have a higher age will achieve a higher adenoma detection rate than an equally experienced endoscopist with a younger patient base. Furthermore, categorisation of adenoma may depend on the attitude of the performing endoscopist [10]. Overall, judging whether one particular procedure is adequate or not remains the responsibility of performing endoscopist.

In light of these difficulties, the ASGE/ACG Task Force on Quality in Endoscopy stated as a key research question, whether the collection of quality indicator data can be automated [44]. Recent years have seen an increasing interest in automatic retrieval of quality characteristics of colonoscopy procedures. Examples are the measurement of insertion and withdrawal time or the detection of landmarks in the intestine, such as the appendiceal orifice, but also the detection of any kinds of lesions. Chapter 2.3 gives a detailed overview of the research in this field.

In the following, some existing medical quality guidelines are reviewed, before describing subjective assessment tools, on which we largely focus our research.

2.2.1 Medical Quality Guidelines

Quality assessment in colonoscopy is the topic of several clinical guidelines, which we summarise in this section. Quality issues are addressed at different levels, from practical guidelines for particular procedures to guidelines for documentation and collection of procedure statistics.

In a guideline document from 1999 [45], the American Society for Gastrointestinal Endoscopy (ASGE) stated recommendations for procedure documentation, listing in detail characteristics of colonoscopy procedures that should be documented. This included, among other things, reporting any findings, complications and the anatomical extent of the procedure. Also included were recommendations for subsequent care.

Another ASGE guideline [46] introduced the collection of procedure statistics that can be compared between endoscopists or endoscopy centres, such as polyp detection rate or the percentage of procedures in which the caecum has been reached. The latter is often referred to as caecal intubation rate.

In Europe, the European Society of Gastrointestinal Endoscopy (ESGE) has released guidelines for image documentation in endoscopy [47]. In the document, first, a number of questions are stated that should be answered for any endoscopic procedure. These include assessing completeness, reasons for incompleteness and morphological descriptions for possibly found lesions. It then lists a total of 8 recommended positions during a colonoscopy procedure, where a still image should be saved for documentation. Beside the ileocaecal valve and the appendiceal orifice, the document also recommends taking still images of the rectum and certain locations in the sigmoid, descending, transverse and ascending colon.

A very recent guideline of the American College of Gastroenterology (ACG) [3] includes, along with documentation and preparation recommendations, explicit descriptions on how the procedure should be performed. The endoscopist is advised to perform a “slow and obsessive examination, designed to expose all of the colonic mucosa” (Appendix 2 of [3]). Also, a mean withdrawal time of at least 6 minutes is recommended to allow for such an inspection.

In summary, it can be said that most of the official recommendations for measuring quality in colonoscopy are long-term statistics of the performance of either endoscopists or endoscopy centres. These measures are useful for continuous quality improvement and review but are less helpful for the decision as to whether a given procedure was adequate. Even the withdrawal time is considered inappropriate for application to individual cases [44]. It was therefore only suggested as an *average* withdrawal time, since the proper individual time depends on the length of the colon and the overall difficulty of the case.

Overall, the assessment of the quality of an individual procedure remains the task of the performing endoscopist. Given that it has been shown that statistical quality metrics such as the adenoma detection rate vary largely among endoscopists [48, 40, 49] and that a relation to mucosal inspection technique is likely [50], it is also likely that self-assessment is an inadequate form for measuring procedure quality. Developing objective quality metrics for individual cases may therefore improve the situation and highlight problems in endoscopic practice. Automatic computation of such metrics may also assist the endoscopist during the procedure or in deciding about a possible full or partial repetition.

2.2.2 Subjective Assessment of Quality in Colonoscopy

An alternative to long term statistical quality measures are subjective assessment tools for colonoscopy. These involve the observation and direct assessment of a procedure by one or more experts, usually guided by a predefined assessment form. A number of organisations have proposed and validated their own protocols for such assessments. We shall outline three validated examples of subjective assessment tools in the following.

- The *Direct Observation of Procedure and Skill* (DOPS) assessment tool [51], developed by the Joint Advisory Group on Gastrointestinal Endoscopy (JAG) in the United Kingdom. JAG DOPS is used within the NHS Bowel Cancer Screening Programme in the UK as an accreditation tool. With 20 criteria, it uses a detailed assessment form that covers a broad spectrum of skills, including assessment, consent, sedation, endoscopic skills and diagnostic abilities.

- The *Mayo Colonoscopy Skills Assessment Tool* (MCSAT) was proposed and validated by Sedlack, et. al, in [52]. The tool was designed to assess trainee endoscopists throughout their training. Therefore, the assessment form includes questions about how much expert assistance was necessary during the procedure. The focus is on cognitive and motor skills, which are encoded using 13 criteria.

- *Global Assessment of Gastrointestinal Endoscopic Skills* (GAGES) was proposed and validated in [12], initiated by the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES). The assessment form is rather general with 5 criteria covering mostly motor skills. Similarly to the MCSAT, it is aimed at trainee endoscopists, and the amount of required assistance is therefore included in most of the descriptors of the criteria.

These tools have their main application in the assessment of the competence of endoscopists. They therefore focus on the endoscopist rather than on the procedure itself. However, apart from the bowel preparation and anatomical particularities of the patient, the quality of the procedure depends on the performance of the endoscopist. The existing assessment tools differ mainly in the amount of detail they capture about the colonoscopy procedure.

Due to their ability to assess many aspects of the quality of individual procedures, we consider the criteria in such assessment tools sensible targets for automatic measurement. The algorithms proposed in this thesis are therefore aimed at measurement of such criteria and are evaluated accordingly.

We concentrate on the JAG DOPS assessment tool, which we consider to be particularly relevant. It is the most mature and in active use within the NHS Bowel Cancer Screening Programme in the UK, while others have yet to be implemented. Furthermore, through our collaborative research activities with Irish hospitals, we have access to clinicians who have participated in a training program for JAG DOPS assessment. This allows us to collect clinically relevant data describing the quality of individual colonoscopy procedures. We describe the development and conduction of the experiment from which we obtained this data in Chapter 5.

JAG Direct Observation of Procedure and Skill

The algorithms presented in this thesis are directed towards automatic measurement of quality criteria used in the JAG DOPS assessment. These criteria form the basis on which we develop and evaluate the mapping from image, video and motion features to procedure quality measures.

The JAG DOPS assessment, as it is used in the UK, has a broader application than we consider in this work. While we seek to assess only the actual colonoscopy procedure, the JAG DOPS assessment includes a number of preprocedure criteria as well as cognitive aspects. In the following discussion of the JAG DOPS assessment we highlight the criteria that fall within the scope of this thesis. The JAG DOPS assessment is divided into four groups of criteria:

1) Assessment, Consent, Communication. This group contains criteria that describe the interaction between the endoscopist and the patient, such as obtaining informed consent and the use of language. These criteria are clearly not measurable from the data we are considering and are therefore out of the scope of this work.

2) Safety and Sedation. Here the ability to use appropriate sedation is assessed, together with the communication with the nursing staff. Again, these criteria are not part of the procedure itself and therefore not dealt with in this thesis.

3) Endoscopic Skills During Insertion and Procedure. This group is the most relevant in the context of our research, as it includes a number of criteria for assessment of the motor skills of the endoscopist. These criteria affect the quality of the procedure and can largely be judged directly from the endoscopic video. They include the quality of luminal views, the usage of steering strategies and control knobs as well as the ability to recognise and resolve loops. We go into more detail about these criteria in Chapter 3. Besides those aspects, the group contains technical pre-procedure tasks such as the checking of the endoscope. The time for completion and awareness of the patient condition are also assessed here.

4) Diagnostic and Therapeutic Ability. This section focusses on the accuracy to which the procedure is carried out. Apart from a number of cognitive criteria, such as the ability to identify landmarks and pathology, the group also contains the assessment of the quality of mucosal visualisation. This is a major quality indicator that we particularly consider, since lesions can only be detected if the area of the mucosal surface is actually visualised.

The full assessment form and the associated grade descriptors can be found in Appendix B. In Chapter 3 we will analyse all DOPS criteria that are especially relevant to our research and discuss possibilities of measuring them automatically. To provide a better context for this, the now following section contains an extensive review of the literature on automatic analysis in colonoscopy.

2.3 Automatic Analysis of Anatomical and Procedural Characteristics in Colonoscopy

Videos from colonoscopy procedures can be described in terms of anatomical and procedural characteristics. Anatomical characteristics include all aspects of the visualised organ, such as:

- the different sections of the large intestine,
- the presence of polyps and other lesions,
- the presence of intestinal contents,
- the locations of landmark points, such as the appendiceal orifice.

Procedural characteristics are, for example,

- the time durations for the insertion and withdrawal phases,
- camera motion,
- image quality,
- the number and type of therapeutic operations, such as taking tissue samples or removing polyps.

Automatic analysis of many of these aspects has been investigated in recent years and this section gives an overview of this field of research.

2.3.1 Detection of Indistinct Frames

Navigation with the endoscope is a demanding task. The length and flexibility of the instrument, and the associated complex force transfer from outside the patient to the tip of the endoscope, leads to an often unpredictable movement of the tip. Therefore, especially during the insertion phase, the camera often comes very close to the intestinal wall or touches it. The lens is also occasionally covered by liquids or other objects. These situations result in very blurry images with no distinguishable anatomical structures or surfaces. Figure 2.3 shows some examples of such images. These images are not suitable for diagnostic purposes. Likewise, the possibilities for automatic measurement of anatomical characteristics are limited. Therefore, these frames are called *non-informative* or *indistinct* in the literature [53, 54].

Problematic in this context is the absence of a universal definition of indistinct frames. For different applications, different degrees of "distinctness" or "clarity" may be useful. This supports the notion of a continuous measure of the clarity of the field of view. However, such a measure has not yet been proposed. We address this problem in Chapter 4 and propose an algorithm for obtaining a continuous clarity measure. In the following we review the literature on the classification problem of detecting indistinct frames in endoscopic video.

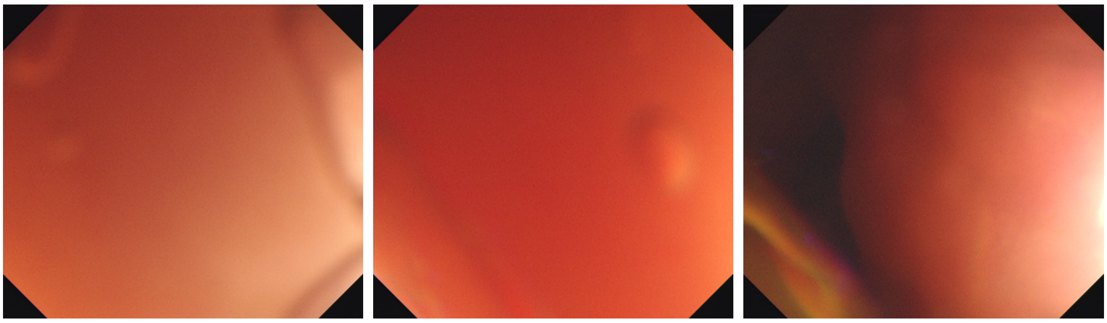


Figure 2.3: Examples of indistinct frames from colonoscopic videos.

The instances in which the camera touches the wall are also referred to as *red-outs*, due to the usual uniform red colouring of these frames. The percentage of time in red-outs is automatically reported by a number of virtual reality colonoscopy simulators as a quality measure. In a study by Grantcharov, et al. [55], an equivalent measure (the percentage of time with a clear view) was found to significantly correlate with the experience of the participating physicians. The number of red-outs was used in [56] as one measure to observe the learning curve of colonoscopy trainees who were training their skills on a simulator. More details on quality metrics used in virtual reality simulators and their validity will follow in Section 2.3.9.

Automatic detection of indistinct frames is motivated by the possible link between their occurrence and the skill level of the performing physician. Apart from that, the average number of such images in a complete colonoscopy video is reported to be between 25% and 37% [57, 53]. Since it is not necessary to search for anatomical or procedural characteristics in indistinct frames, their detection can be seen as a valuable preprocessing step to a more sophisticated analysis system for colonoscopic videos. The benefits are faster computation and a more homogeneous dataset for the following computational steps.

Indistinct images are usually defined as very blurry, which is equivalent to the absence of rapid intensity changes or an energy concentration in the lower spatial frequency bands. The different approaches reported so far all make use of this property. There are, however, certain indistinct images where this property is less prominent, as they contain sharp edges due to specular highlights or air bubble contours. Then again, images can be clear and in focus but show very uniform mucosal surface and therefore have less high frequency content. Figure 2.4 shows examples of these types of images. For a high detection accuracy, these exceptions need to be taken into account.

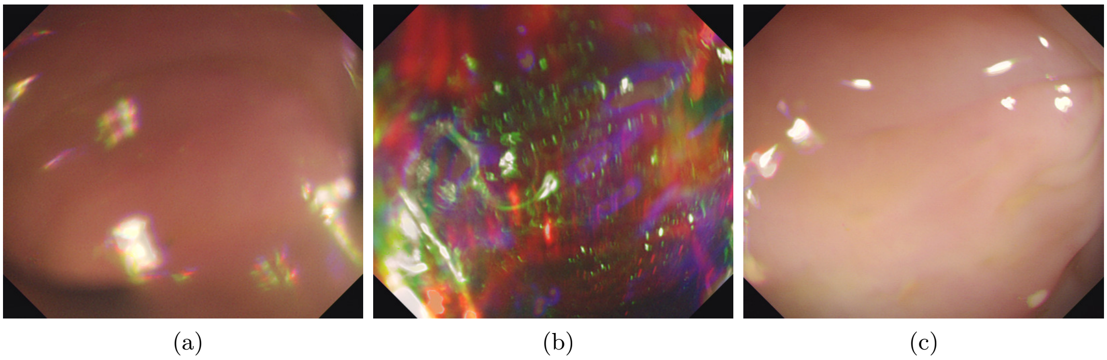


Figure 2.4: Examples of problematic images in the context of indistinct frame detection. (a),(b) indistinct frames showing specular highlights and air bubble contours; (c) informative frame showing a smooth part of the colonic mucosa.

In [58], Cao, et al., proposed an approach to scene segmentation for colonoscopic videos, based on the detection of indistinct frames. In order to detect those, the authors used the standard deviation of the spatially reduced image representation that can be directly obtained from MPEG-2 compressed video in the form of the biases of a blockwise discrete cosine transform of the images (see, e.g., [59]). This approach therefore achieves a high computational efficiency for MPEG-2 compressed video. However, while the approach may be sufficient for the purpose of scene segmentation, indistinct frames with high contrast and informative frames with low contrast may be misclassified.

An edge based approach was proposed by Oh, et al., in [60]. After applying a Canny operator [61] and thresholding the image, the edge pixels were split into two categories: Isolated edge pixels with no edge pixels in their neighbourhood, and connected edge pixels with at least one edge pixel in their neighbourhood. The authors defined an *isolated pixel ratio* as the percentage of isolated edge pixels among all the edge pixels and used it as the discriminating feature.

In [57], An, et al., proposed an unsupervised frequency domain approach to detecting indistinct frames. It was based on the observation, that the magnitude of the discrete Fourier transform (DFT) of an indistinct frame shows textural characteristics different from those of the DFT magnitude of an informative frame. It used texture features of the DFT representation and k-means clustering to discriminate between informative and indistinct images. Without incorporating prior knowledge into the clustering algorithm, it is highly sensitive to varying distributions of the feature vectors for different videos. However, no constraints on the k-means algorithm were reported in the paper.

The approaches presented in [60](edge-based) and [57](GLCM) were compared in [54] and complemented by a method for detecting and removing specular highlights. As explained earlier, specular highlights can lead to classification errors. The specular highlights were detected in HSV colour space in two steps. First, pixels with value $> T_v$ and saturation $< T_s$ were defined as specular highlights, with T_v and T_s being fixed thresholds on value and saturation, respectively. Then, a segmentation of the image was performed using JSEG [62], which segments the image into regions according to colour and texture homogeneity properties. Specular highlights that were not recognised in the first step were found using outlier detection in each detected region. It was reported that, without specular highlight detection, the approach from [60] achieved an accuracy of 91%, while the approach from [57] achieved 95%. Prior specular highlight removal led to improvements for both methods: An accuracy of 95% was assessed for the edge-based and 97% for the GLCM approach.

The reported performance of the existing approaches to indistinct frame detection is impressive. However, following an implementation of the GLCM approach, we could not reproduce similar results. Details on this issue can be found in [53], where we propose an approach to indistinct frame detection based on wavelet decomposition and evaluate our method against the GLCM approach in [60]. In Chapter 4 we present an approach that extends our previous work by measuring the clarity of the field of view on a continuous scale instead of the binary classification into informative and indistinct frames.

2.3.2 Detection, Localisation and Segmentation of the Colon Lumen

The term *lumen* generally refers to the inside space of a tubular structure. During a colonoscopy, when navigating through this inside space, the lumen is not an object as such, but rather the most distant distinguishable region in the image, when the camera has a clear view aligned with the colon. Figure 2.5 shows some examples of images showing the colon lumen. The lumen is an important feature for the endoscopist, especially during the insertion phase, as it determines the direction in which the endoscope should be advanced. For patient safety reasons, it is recommended to only advance the endoscope when being certain about the lumen location [63]. For screening purposes the lumen is an important reference point, since all mucosa surrounding the lumen has to be visualised. The ability to maintain a luminal view throughout the procedure is therefore a recognised major quality criterion. This section reviews approaches towards detection, localisation and segmentation of the colon lumen. A significant number of methods for these tasks have been reported to date. However, the problem of assessing the quality of luminal views and the ability to maintain luminal views throughout the procedure has not been previously addressed. In section 4.4 we propose an approach to this problem.

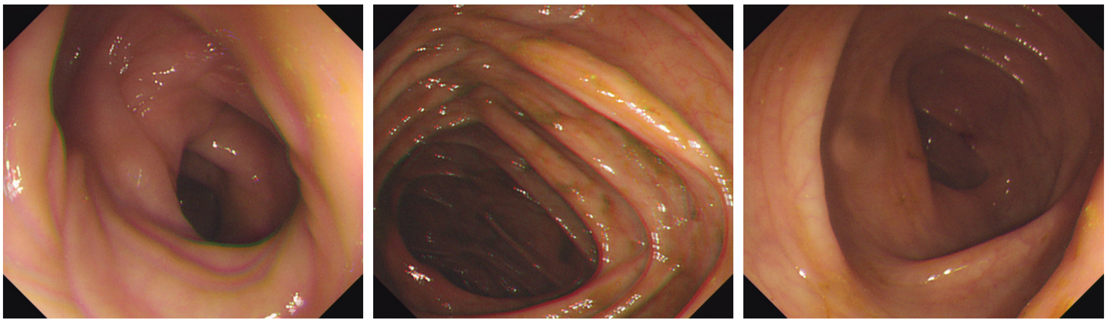


Figure 2.5: Examples of colonoscopic video frames showing the intestinal lumen.

The lumen in endoscopic images is usually the region that is most distant to the camera and therefore also most distant to the light source. For a clean mucosa and uniform lighting this makes the lumen region the darkest region in the image. Consequently, this property is usually exploited in approaches to lumen detection. However, a number of difficulties arise when an algorithm relies on this property alone. As can be seen in the example images in Fig. 2.5, the brightness of the lumen region can vary largely between images. Apart from that, *diverticula*, pathological pouches in the mucosa that develop in some colons, are easily mistaken for the colon lumen. Problems also occur in case large dark objects are present in the intestine, such as faecal materials. Therefore, it is necessary to take additional properties of the lumen into consideration, such as its shape. Moreover, searching for the darkest region in the image does not answer the question, whether the lumen is visible in the image or not. Since the lumen can appear in various shapes and sizes, the solution to this problem is not straightforward. However, knowing whether the lumen is present or not may be valuable information for procedure analysis, e.g., discriminating between global and close mucosal inspection as proposed in [64, 65, 66]. Furthermore, in automatic navigation applications, steering towards a suspected lumen region in an image that does not show the lumen, may even be harmful to the patient.

Nevertheless, the majority of approaches concentrate on the segmentation of the lumen and often focus on obtaining an accurate lumen boundary. Since the lumen is not an object as such, the exact boundary location is a rather subjective property. Different observers might see the lumen region begin at different depths. Consequently, speaking of an accurate segmentation of the lumen region is not necessarily helpful.

Khan [67, 68] proposed an approach to detect the dark region corresponding to the intestinal lumen using an iterative search algorithm in a quad-tree representation of the image. Their method was to find the largest quadratic region, in the quad tree, that is uniform and has an average intensity close to the first peak in the grey level histogram of the image. Having found this region, the outline was refined using the finer levels of the quad tree. In [69], Sucar and Gillies combined this approach with a classification system to decide whether the detected region really corresponds to the intestinal lumen. They used the location and the size of the region in question as well as intensity statistics in order to make that judgement and combined these features with information from previous and following frames. Both approaches assume that the lumen is a uniform and dark region in the image. Given the image quality and resolution of current endoscopy systems, there is often a considerable amount of structure in the lumen region. Uniformity is therefore no longer a valid assumption.

Sucar, et al. [70], estimated the lumen position using a shape from shading approach described in [71]. Using the assumptions that the light source is a point source located at the camera position and that the surfaces are lambertian with a slowly varying reflection coefficient, the method computed the surface gradients for each pixel given its intensity and its intensity gradient in two orthogonal directions. The lumen location was then approximated from a histogram representation of those surface gradients. The authors reported that the method was error prone when the lumen was located close to the centre of the image.

In [72], Kwoh and Gillies proposed to use Fourier domain information for finding the location and size of the lumen in an image. Their approach started with a calculation of one dimensional discrete Fourier transforms (DFT) of the cumulative intensities along the x and y directions, followed by a template matching step in the Fourier domain. The template matching procedure returned approximate position coordinates and size of the lumen. Interestingly, the method could give an indication of the lumen position, even when the lumen was located outside of the image, which makes it particularly useful for navigation applications. The authors demonstrated high correlation between the predictions of the method and expert labelling of lumen position and size. Kwoh, et al. [73], later integrated the approaches from [67], [70] and [72] into a Bayesian network in order to combine the strengths of the different methods. The evaluation outlined in the paper unfortunately lacks the necessary depth and remains inconclusive as to how the presented results were obtained.

Krishnan, et al. [74], used an adaptive thresholding approach proposed by Tsai and Chen in [75] to obtain a binary image. The centroid of the dark pixels in the binary image was chosen as seed point for a radial region growing algorithm operating on homogeneity and edge information from the original image. The method inherently leads to inaccurate lumen positions in the presence of any dark regions other than the lumen, as they drag the centroid away from the centre of the lumen. It may even drift outside of the actual lumen region, leading to a non-lumen seed point for the region growing algorithm.

Similarly, Kumar, et al. [76] used adaptive progressive thresholding based on Cheriet's method [77]. Again, the centroid of the obtained region was used as seed point for a differential region growing algorithm to obtain an accurate lumen boundary. Later, in [78, 79], the same authors proposed a real time modification of the approach replacing the region growing algorithm with an integrated neighbourhood search approach and representing the image by a quad structure. Similar approaches based on thresholding and region growing were proposed in [80, 81, 82, 83, 84, 85]. These methods have in common that that they assume the lumen to be the darkest region in the image. This simplification results in suboptimal performance on less ideal examples.

In [86], Krishnan, et al., performed a region segmentation of the image following the histogram based approach in [87]. Using a number of features from these regions, a fuzzy rule base was constructed to discriminate between lumen, polyps, bleeding lesions and background. A region was labelled as lumen, if its area was *middle*, and its mean value of intensity and saturation was *low*. The authors have not reported an evaluation of the method.

Hwang, et al. [64], segmented the image into regions of similar colour and texture using the JSEG image segmentation algorithm presented in [62], which involves colour quantisation and region growing and merging at multiple scales. Convexity, size and intensity were then used as criteria to identify the lumen region. The objective of the authors was to discriminate *lumen views* (images containing the lumen) from *wall views* (images showing only the intestinal wall). With a similar objective, Cao, et al. [88], calculated a likelihood $\mathrm{P_{NoLumen}}$ of the absence of the lumen:

$$\mathrm{P_{NoLumen}} = \frac{\mathrm{I_{DarkestRegion}}}{\mathrm{I_{Max}}},$$

obtained from the ratio of the mean intensity of the darkest region in the image, $\mathrm{I_{DarkestRegion}}$ (following JSEG segmentation), to the maximal intensity in the image, $\mathrm{I_{Max}}$. A problem of these approaches is the computationally expensive JSEG algorithm. Both algorithms also make quite strict assumptions on the lumen region, either assuming convexity of the lumen region, or assuming the lumen region must be the darkest region in the image.

Discrimination between lumen views and wall views was also addressed by Liu, et al., in [65]. In their approach they obtained a pixel level classification into lumen and wall pixels using a decision tree classifier and the colour channel intensities of single pixels as feature vectors (various colour spaces were compared). Global features of the resulting binary image and the size and position of its foreground objects were then used for image classification. The method was designed to work on clear images, as all blurry images were removed from the data set. In a very recent publication [89], the same research group proposed a novel method that uses the characteristic pattern of folds in the colon to detect the lumen position. The folds are extracted from an edge image and arranged into “quasi-parallel” groups from which the lumen position is inferred. Unfortunately, there is no information on how the edge image was calculated. Similarly to their earlier method, it is designed to detect the location of the lumen only in clear images, where the edge detection returns reliable results.

In the context of wireless capsule endoscopy, Zabulis, et al. [90], used a *mean shift* algorithm [91] in order to find the regions with lowest intensity and highest intensity in the image. The low intensity regions were assumed to correspond to the lumen, while the high intensity regions were associated with *highlights*, i.e., tissue in the vicinity of the lens that appears very bright. The authors argued that the locations of highlights were indicating the camera pose and proposed them as a cue for navigation applications. The mean shift algorithm was initialised with multiple seed points at different locations in the image. The points where the algorithm converges, were expected to form clusters at the lumen and highlight positions. Since other clusters may form as well, each cluster was represented by a region, that was evaluated for its area, compactness and mean intensity and pruned accordingly. The regions were found using an intensity based region growing algorithm followed by a contour smoothing step. When the mean shift algorithm converged at points that were spread widely over the image or were located at the periphery, the lumen was assumed to be absent.

As can be seen from this overview on lumen detection and segmentation algorithms, the majority of approaches used local brightness as the main discriminating feature. The darkest region in the image was either directly assumed to be the intestinal lumen, or after checking additional properties such as the region’s shape and size.

In Section 4.4 we report a method that integrates the detection of the presence and position of the intestinal lumen in the image with the assessment of the quality of the luminal view. Our method is based on a region candidate selection using intensity, colour and shape features and support vector machines.

2.3.3 Detection of Lesions

In colorectal cancer screening, the prominent lesions are colorectal polyps. However, various other pathological findings can be observed during a colonoscopy procedure. Those include, among others, *diverticular disease*, *colitis*, *lower gastrointestinal bleeding* and *vascular malformations* [92]. The majority of research in automatic lesion detection from colonoscopic videos deals with polyp and cancer detection, or general classification of images into normal and abnormal. In a typical automated colon abnormality detection system the high-level characteristics of colon cancer, as discussed above, are translated into image-based features. Following this feature extraction, a discriminant analysis is performed using a training and a test dataset. For an overview on lesion detection in endoscopy, see, e.g., [93, 94].

One major source of motivation for automated lesion detection is the recently reported miss rate in colonoscopy procedures [31, 32, 33]. In order for automated detection methods to improve the situation, a significant percentage of these missed lesions must be within the camera's field of view during the procedure. Whether this is the case, has, to the best of our knowledge, not yet been sufficiently investigated. It is therefore also possible that a high percentage of missed lesions are missed due to poor bowel preparation or incomplete visualisation of the mucosal surface. The benefit of automated lesion detection may consequently be limited.

It is beyond the scope of this research programme to analyse this issue. However, our discussions with expert gastroenterologists suggested that lesions in the field of view of the camera are generally detected by the endoscopists. Because of the mentioned open questions, this work focusses on measuring quality issues that may cause the high miss rates.

2.3.4 Detection of Intestinal Contents

Prior to the actual colonoscopy procedure, the large intestine needs to be cleared of any intestinal contents, i.e., faecal materials and undigested food. This process is called *bowel preparation*. A variety of preparation methods exist, the most popular being diet and cathartic regimens, gut lavage and phosphate preparations [95]. Unfortunately, there is no optimal method. Good cleansing performance is often complemented by larger patient discomfort and side effects. A comprehensive discussion of preparation methods and recent recommendations can be found in [96].

According to [95], poor preparation can lead to improper visualisation of the mucosal surface. Lesions can be missed [36] and it is more likely that the procedure can not be completed. In addition, perforation of the intestinal wall, one of the possible complications during a colonoscopy procedure, is more likely to result in dangerous septic complications when the bowel preparation was inadequate.

Currently, no objective measure of preparation quality exists. Methods for bowel preparation are usually compared in the literature using a semi-quantitative scale system where experts rate procedures according to their subjective opinion [97]. This is either an overall rating of the procedure (e.g., the *Aronchick* scale [98], a five point scale with associated grade descriptors) or a more detailed rating of parts of the large intestine (e.g., the *Ottawa* scale [99, 100], a four point scale for each of the three main segments of the large intestine).

Automatic detection of intestinal contents and measurement of the impact on the visualisation of the mucosal surface may be helpful both for clinical research and practice, as it can potentially offer a higher objectivity compared to manual measures.

Little research into intestinal contents detection has been done to date. Hwang, et al. [101], investigated raw colour vectors and colour histograms as features for stool detection in colonoscopy images. They used a support vector machine classifier (see, e.g., [102]) and compared several colour spaces. The HSV colour space was found to yield optimal results for distinguishing images containing stool from clean images. Vilariño, et al. [103], used a bank of Gabor filters (see, e.g., [104]) to detect intestinal juices in images from wireless capsule endoscopy. The filter outputs are summed up, yielding a large response for the characteristic structure of intestinal juices, i.e., a texture made up of small bubbles. Unfortunately, in conventional video endoscopy, not all intestinal contents have similar texture properties.

To the best of our knowledge, no approaches have been published that aim towards quantifying the intestinal contents in a way that can be linked to a measure of preparation efficacy. Because of the mentioned links to procedure quality, such an objective measure would be highly desirable. In this thesis, we evaluate quality criteria in a clean, simulated environment, where issues of bowel preparation do not apply (see Chapter 5). We therefore leave the measurement of preparation efficacy to future research.

2.3.5 Detection of Interventions

Endoscopes have a channel through which the insertion of specialised medical instruments is possible. It is common to take tissue samples during endoscopy procedures. Additionally, certain diseases or lesions can be treated directly. An example is the removal of polyps in colonoscopy. This intervention is called *polypectomy* and can prevent the development of malignant tumours.

Various instruments are used during these interventions. Examples are biopsy forceps or injection needles [7]. Figure 2.6 shows examples of interventions in colonoscopy procedures. The detection of such interventions can be valuable for summarisation and documentation of endoscopy procedures. Scenes containing an intervention could then easily be reviewed for quality control purposes or in the follow-up surveillance of the lesions.

The problem of detecting interventions can be addressed by detecting medical instruments. While the tracking of medical instruments has been studied in the context of laparoscopic surgery [105, 106, 107], little research of this kind has been done for application in endoscopy screening.

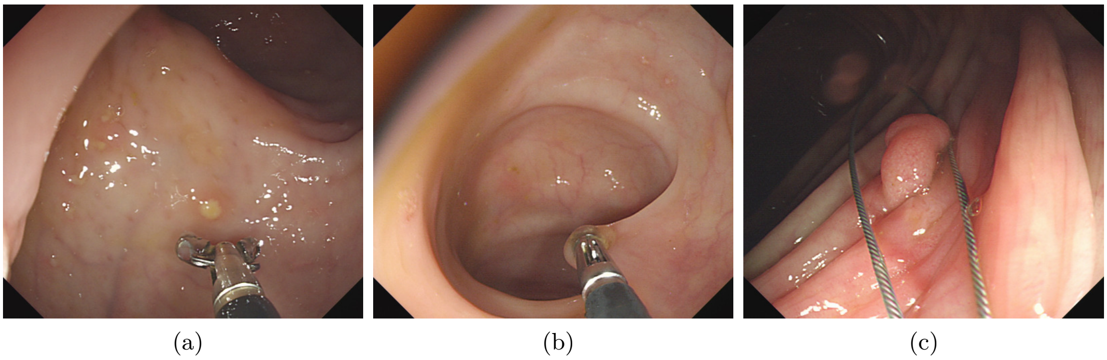


Figure 2.6: Examples of interventions during colonoscopy procedures: (a) Biopsy forceps about to collect a tissue sample, (b) Polyp removal using a biopsy forceps, (c) Polyp removal using a snare.

Cao, et al. [108], proposed a region based approach to medical instrument detection in colonoscopy videos. They first segmented the image using JSEG [62], followed by a position dependend filtering of the resulting regions, as instruments can only appear in certain areas of the image. Possible over-segmentation was tackled using a shape based region merging algorithm. The resulting regions were then matched against a template database of instrument wire regions using Fourier shape descriptors (see, e.g., [109]). Images containing instruments were combined to *operation shots*, i.e., series of consecutive images containing a medical instrument.

In [110], the same authors proposed an enhancement of this method by using a texture based region growing algorithm instead of the region merging step. The region matching was performed using moment invariants [111] instead of Fourier shape descriptors.

2.3.6 Detection of Caecal Landmarks

A colonoscopy procedure is considered *complete* when the caecum has been intubated [7], i.e., the endoscope has been advanced to the full extent of the large intestine. It is common practice to document this by taking still images of the visible landmarks inside the caecum. Examples of such landmarks are the *ileocaecal valve* (the junction to the terminal ileum) or the *appendiceal orifice* (the junction to the appendix). Intubation of the terminal ileum through the ileocaecal valve is also considered helpful for documentation. In the terminal ileum, the mucosal surface shows a granular texture due to the *villi*, which are small, finger-like, protruding structures. The *crow's foot* or *y-fold* is another useful, but less reliable landmark. Figure 2.7 shows examples of some of the caecal landmarks mentioned here.

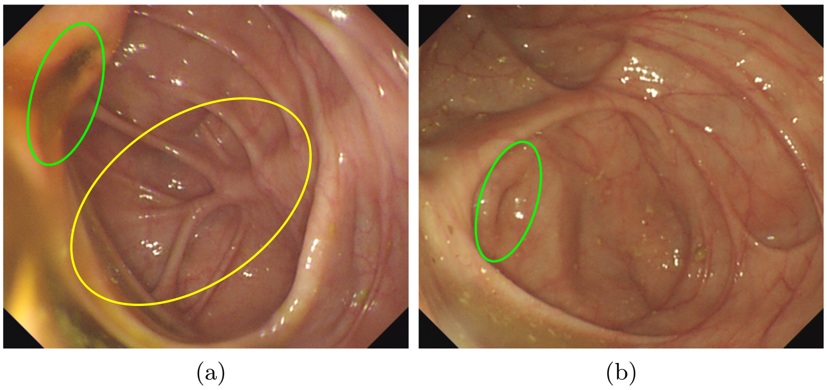


Figure 2.7: Examples of caecal landmarks: (a) Ileocaecal valve (green) and crow's foot (yellow), (b) appendiceal orifice.

Detection of the ileocaecal valve has so far only been addressed in virtual colonoscopy [112, 113], where it is likely to be confused with a polyp, increasing the number of false positives in computer aided diagnosis systems. As mentioned before, techniques from virtual colonoscopy are usually not directly applicable to video colonoscopy and are therefore not reviewed in detail in this report. In the context of video colonoscopy, only the detection of the appendiceal orifice has been addressed so far. Cao, et al. [88], derived a number of features relating to properties of images containing the appendiceal orifice. The properties they used were the assumed absence of the intestinal lumen and the position and shape of the appendiceal orifice. The orifice was assumed to be in the centre of the image and to show contours in the shape of segments of ellipses. The first property was expressed by a likelihood of absence of the lumen described earlier in section 2.3.2. The other properties were described by attributes of contours that have a certain curvature or can be considered segments of ellipses. The assumptions made for this algorithm only hold for a certain range of viewing angles, since the contours become less similar to segments of ellipses when outside a certain range. Also, some of the features used are highly scale and illumination dependent.

Wang, at al. [114], proposed a more careful contour selection step, discarding contours according to their length and curvature, and also those that were found to belong to the edges of specular highlights. The remaining contours were classified using a number of features obtained from the edge cross-sections - the neighbourhood on a normal line (perpendicular to the tangent) of the contour at a given pixel on the contour. Examples of such features are the number of edge crossings along the neighbourhood line or the difference of the saturation on both sides of the contour. Similar methods have been used, e.g., for the recognition of vessels in biomedical images (see, e.g., [115, 116]). In [117], the same authors propose an enhancement of their algorithm by allowing only images that are sufficiently similar to neighbouring images in the video, according to a block-based histogram distance measure, which is assumed to correlate with the amount of camera motion.

Given that caecal intubation is routinely documented by taking a still image of caecal landmarks, the benefit of an automatic detection of caecal landmarks is limited. Especially, since it has been shown that even experts disagree on whether certain landmarks are visible in images [118, 119]. This fact also limits the quality of obtainable ground truth data and thus raises questions about the validity of the evaluation of developed approaches in this field.

2.3.7 Camera Motion Estimation and 3D Reconstruction

Camera motion estimation can be used to make judgements about semantic aspects of colonoscopy procedures and the quality of the inspection. For example, very fast camera movements indicate a less thorough examination of the mucosal surface, while a very still camera might suggest that an object in the field of view is being closely examined.

A motivation of 3D reconstruction approaches in colonoscopy is the prospect of measuring a potentially very accurate indicator of procedure quality: the percentage of visualised mucosal surface. Another possible application is the generation of an accurate model of the large intestine, comparable to the data obtained from virtual colonoscopy, that might allow for an efficient offline screening of the large intestine. 3D information can also be used to register visual image data to preoperative data from other image modalities, such as computed tomography (CT) or magnetic resonance imaging (MRI). Successful registration would allow for a better guided search for lesions that were found in the preoperative 3D data and are to be treated or removed.

The fact that the large intestine is not a rigid body but rather a highly deformable structure increases the difficulty of both camera motion estimation and 3D reconstruction significantly. Other problems can arise from specular highlights and liquid intestinal contents. Both can move inconsistently compared to the actual structure of the currently visualised scene. Furthermore, obtaining a complete and continuous model of the large intestine is often impeded by long indistinct video segments. After such a segment, the camera may find itself in a completely unknown environment unable to find any structural correspondences.

Despite the variety of applications in the analysis of endoscopy procedures, we can not consider camera motion estimation or 3D reconstruction in this thesis. The research presented here is embedded in a broader research project, where these topics are addressed by others. The algorithms in this branch of the research project are not yet developed to a state where they could be applied in the context of this work. Therefore, we focus on methods that do not incorporate camera motion estimation or 3D reconstruction.

2.3.8 Endoscopic Video Segmentation and Summarisation

Automatic semantic segmentation and summarisation of endoscopic videos has many useful applications. An obvious example is in medical video databases, in which the stored videos could be reviewed and searched more efficiently through a segmentation into meaningful scenes. These scenes could relate to anatomical segments of the colon, but could also reflect procedural segments, such as the insertion and withdrawal phase, or events, such as interventions or the visualisation of lesions or landmarks. This section summarises previous research into automatic temporal segmentation of colonoscopic videos and the inclusion of detected events in summarisation frameworks. Certain algorithms reviewed here make use of some of the detection and segmentation algorithms described earlier in this chapter.

Cao, et al. [108, 120, 58], used speech recognition to obtain a semantic segmentation of colonoscopic videos. In their study, the endoscopists used predefined phrases when entering and leaving anatomical segments of the large intestine, complemented by certain terms such as *polyp* or *cancer* to mark abnormalities during the procedure. When a certain segment was not recognised by the speech detection algorithm, the visual data was searched for a certain pattern, namely a sequence of clear frames followed by blurry frames followed by again clear frames, which the authors reported being a common pattern observed at transitions between the segments. A problem with this approach is that having to name segments and events during the procedure increases the cognitive load of the endoscopists, which may have a negative effect on the procedure outcome.

In [64], Hwang, et al., proposed to accumulate the dolling (forward and backward) motion of the camera throughout the video and detect the transition between the insertion phase and the withdrawal phase by picking the frame number that corresponds to the highest accumulated motion value. This approach was also taken in [121]. These methods use data from the whole video and thus can only be used off-line. To overcome this problem, Oh, et al. [122], proposed to declare the current maximum of the accumulated motion during the procedure as a temporary phase transition, giving a final confirmation at the end of the video signal reception. A problem with these methods is the general assumption that the real camera dolling motion is well approximated by the motion estimation algorithm. This algorithm, however, needs reasonably clear images in order to achieve this. As mentioned before, many images in colonoscopic videos are extremely blurry, leaving the actual camera motion unclear to the algorithm. The endoscope is usually repeatedly moved back and forth during both the insertion and the withdrawal phase, and it is possible that the accumulated distance the camera travels is much larger than the actual length of the colon. Errors of the motion estimation can therefore have a big impact on the detected phase transition, especially when the errors tend to be larger in one of the two possible directions. In Chapter 6, we propose an alternative approach which makes use of endoscope motion measured at the orifice.

Stanek, et al. [123], took a colour based approach to discriminating images taken outside the patient's body from images of the actual procedure, in order to detect the start and end of the procedure. The features they used were the intensity of the red channel both independently and in relation to the overall brightness of the image. The authors demonstrated the reliability of the method, making it a viable alternative to manual video editing for archiving purposes.

In the context of wireless capsule endoscopy (WCE), Iakovidis, Tsevas, et al. [124, 125], used an unsupervised method for summarising a capsule endoscopy video by representative images. They used an approach reported by Okun and Priisalu in [126], which may also be directly applicable to colonoscopy videos. In a first step of the approach, the video frames, represented by a grey-scale vector feature, were grouped into clusters using a fuzzy c-means clustering algorithm [127]. Then, a non-negative matrix factorisation [128] was carried out on the clustering result. The representative frames for each cluster were determined by applying orthogonality constraints. In their evaluation of the method, none of the abnormalities in their test data set were missed in the video summary while the number of frames to view was reduced significantly. Such an approach may be useful for automatic generation of reports in colonoscopy. Abnormalities in colonoscopy, however, differ from what can be found in WCE, and the method may be less successful in this different domain.

Some other works on video segmentation for wireless capsule endoscopy have been published (e.g., [129, 130]) that are too domain specific to be directly applicable to conventional endoscopy, since they deal with detecting transitions between the different organs that are examined in the procedure.

Generally, most of the algorithms reviewed in this chapter allow for a temporal segmentation of endoscopic videos. By labelling images with the properties that were automatically determined, parts of the videos can be made accessible more easily. For example, one could just look at images where the intestinal lumen is visible or at the part of the video that shows therapeutic interventions. Cao, et al. [108], for example, implemented a browsing tool that allows the user to access all the segmented scenes of the colonoscopic video using a representative image for each scene.

By collecting all obtainable information about the videos in a central framework and putting it into a common context, it may be possible to generate a detailed summary of the colonoscopy procedure, including relevant quality metrics. Such information, combined with an interlinked video browsing environment may support the clinical analysis and decision making process and improve the overall quality of service.

2.3.9 Automatically Determined Quality Measures

In a document of the ASGE/ACG (American College of Gastroenterology) taskforce on quality in endoscopy [44], it was stated as a key research question, whether the collection of intraprocedural quality indicator data can be automated. Automatic computation of quality indicators from colonoscopy video data has been only marginally addressed in the literature, despite the interest of the medical research community. In this section, we give an overview of the literature on this topic.

Hwang, et. al [64], combined indistinct frame detection, camera motion estimation and lumen recognition to obtain a number of quality measures that are mostly related to durations of semantic segments of the video. Examples are the duration of the insertion phase and the withdrawal phase, or the duration of the withdrawal phase disregarding all indistinct frames (the *clear withdrawal time*). An interesting measure is also the ratio of *wall views* and *lumen views*, which should be properly balanced, according to the authors. In a more recent article [121], the same authors included also their intervention detection method to determine the clear and intervention-free withdrawal time.

Liu, et al. [66], addressed a different aspect of procedure quality by trying to evaluate, whether the camera was pointed at all sides of a colon segment. They defined the location of the lumen as the centre, subdivided the view deviations from the lumen direction in four quadrants and computed a histogram of the number of images, in which the camera was pointed in the direction of the different quadrants. The authors argued that examination of all four quadrants is desirable. This can be seen as a first attempt to measure the amount of mucosal surface that was visualised during a procedure. Liu, et al., recently proposed an amended version of their approach [89], mainly with an improved method for detecting the lumen position. Apart from this, the histogram measure was replaced by counting the number of spirals (the coverage of all 4 quadrants) in a procedure. Both approaches do not take into account the forward and backward motion of the camera. A fast movement through a 20 cm long segment of the colon would get a similar score as a careful slow inspection of a very short segment. Apart from that, the authors assume that the endoscopist examines what is located in the centre of the endoscopic field of view. However, pointing the tip of the endoscope in exactly the desired direction is difficult. Endoscopists may have to accept suboptimal camera positions frequently. It is

Table 2.1: List of quality measures proposed in the literature.

Measure	Details
Insertion time (IT, [64])	Duration of the insertion phase
Withdrawal time (WT, [64])	Duration of the withdrawal phase
Clear withdrawal time (CWT, [64])	WT minus the duration of indistinct shots
Clear withdrawal ratio (CWR, [64])	CWT/WT
Number of camera motion changes (NCMC, [64])	Number of forward/backward motion changes during the withdrawal phase
Ratio of camera motion changes (RCMC, [64])	$NCMC/CWT$
Wall-lumen inspection ratio [64]	Number of wall view images divided by the number of lumen view images
Wall inspection fraction [64]	Number of wall view images divided by the number of indistinct images
Clear operation-free withdrawal time [121]	CWT minus the duration of operation shots
Average quadrant coverage score [66]	Average number of quadrants visualised in N consecutive lumen views
Spiral number [89]	Number of spirals performed during a given duration of a procedure

therefore likely that the assumption of a central region of interest is not entirely valid. Such perceptual issues may be further investigated by analysing the gaze positions of endoscopists while they perform a procedure on a colonoscopy simulator. Studies in this direction have already been reported in [131] and [43].

In summary, the literature offers only few suggestions for quality measures, that can be determined automatically. Table 2.1 lists the measures proposed in the literature so far. The insertion and withdrawal times are, as pointed out in Sect. 2.2.1, only valid as quality measures when looking at an average over many procedures. Measures for individual cases, such as the wall-view/lumen-view ratio or the quadrant coverage histogram, have yet to be evaluated for validity. Because of the problems that were pointed out here, a more accurate measure of mucosa coverage is definitely desirable. We consider it beneficial to look into measures that represent generally accepted insertion and examination techniques and best practices, which is the chosen direction for this thesis.

Score Systems in Virtual Reality Colonoscopy Simulators

An application, where performance is already being automatically rated for individual cases, is in scoring systems for virtual reality colonoscopy simulators, which are used in the training of endoscopists. In such simulators, the endoscope is maneuvered through a virtual 3D model of a colon [132, 133, 134]. Current simulators offer, apart from the visual simulation, haptic feedback for a realistic feel of the bends of the colon or when touching the mucosal surface. Furthermore, audio feedback is implemented to give a sense of patient discomfort during the procedure.

Virtual reality simulators have the advantage, that both the camera position and the environment is known. This facilitates the automatic measurement of quality indicators significantly. Consequently, a number of such measures are readily implemented in these simulators. The model commercially available from Simbionix Ltd., Israel, for example, measures the total duration of the procedure, the percentage of visualised mucosa and the time spent in *red-out*, among others. The term *red-out* refers to the, usually intensely red, indistinct frames occurring when the camera tip touches the mucosa. The simulator also gives an estimate of patient discomfort caused by the procedure. Unfortunately, the heavy use of knowledge about the simulated environment makes virtually none of the simulator methods usable for the assessment of real colonoscopy procedures.

Nevertheless can research efforts for the measurement of real colonoscopy procedures benefit from numerous studies that are constantly carried out to validate simulators. It has been shown, for example, that the mentioned simulator can reliably discriminate between novices and experts [135, 136, 137]. Studies like these can give valuable input to the search for quality metrics for real colonoscopy procedures.

2.3.10 Summary of Literature Review

In this chapter, we have discussed literature on the topic of quality in colonoscopy and approaches to measuring quality related characteristics of colonoscopy procedures automatically. There exist a number of approaches for measuring single characteristics such as the lumen position or the presence of intestinal contents. However, only a few publications contain attempts to assess the quality of colonoscopy procedures itself. These approaches mostly lead to the proposition of novel quality measures, which are rarely assessed for their clinical relevance.

In this thesis we follow a different route by focusing on existing, validated quality measures for individual procedures, and select relevant characteristics based on the objective of measuring these automatically. The following chapter begins this process by discussing these target measures and their underlying procedure characteristics.

Chapter 3

Data Sources and Target Quality Measures

The previous chapter has provided an overview of existing approaches to measuring various aspects of colonoscopy procedures. In the following we will introduce the sources of data we have available and identify the characteristics we consider to be most relevant for our objective to assess quality in colonoscopy according to JAG DOPS assessment criteria.

3.1 Data Sources and Preprocessing

For the measurement of quality related characteristics, we have two sources of data available. One is the visual information from the camera of the endoscope in the form of high definition video data. The video processing unit of the used endoscopy system outputs 25 interlaced frames per second with a size of 1920 px × 1080 px. The actual endoscopic image is smaller than the full frame, usually 1076 px × 928 px, with black triangles in the corners due to the octagonal sensor design. The endoscopy system uses sequential RGB image acquisition. In this method the colour information is captured by one monochromatic photo sensor at different time instances under sequential red, green and blue illumination. This allows higher resolution per sensor area at the expense of colour misalignment artefacts, which occur during fast camera movement (see Figure 3.1 for some examples).

The system maps the acquired colour channels to the output video frames using a pulldown pattern[1]. All 3 colour channels are acquired at a rate of 20 Hz. This has to be mapped to an output of 25 interlaced colour frames per second. In general, an interlaced

[1] The term *pulldown* refers to the use of video interlacing to convert video data between different frame rates.

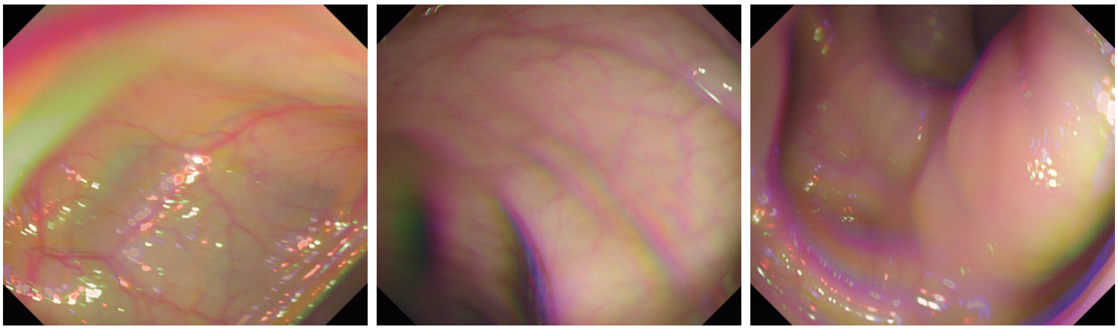


Figure 3.1: Examples of images with colour channel misalignment.

frame consists of two *fields* with the same width but half the number of pixels on the vertical axis. To form a frame, the two fields are interleaved by alternating between them for each horizontal line in the frame (the field mapped to odd lines is commonly called the *top* field, while the one mapped to even lines is called the *bottom* field). The actual mapping to be performed is therefore from the acquired 20 colour images per second to 50 video *fields* per second. This is done in a way that with each consecutive field, the single colour channels are replaced according to a recurring pattern. We denote the colour channels as R_i, G_i and B_i, with the subscript i corresponding to the place in the sequence of acquisition. Assuming *top-field-first* interlacing, the colour channel mapping follows the following pattern (T and B stand for *top field* and *bottom field* in the illustration):

Frame	1		2		3		...
Field	T	B	T	B	T	B	...
Red	R_1	R_1	R_1	R_2	R_2	R_3	...
Green	G_1	G_2	G_2	G_2	G_3	G_3	...
Blue	B_1	B_1	B_2	B_2	B_2	B_3	...

This means that a full change of all colour channels is completed once after 3 field changes and once after 2 field changes in an alternating manner.

Given this knowledge we preprocessed the video data as follows. We separated the fields and interpolated them to the size of the full frame using the edge-based line average algorithm [138]. In the resulting 50 fps video stream the colour channel pattern is detected by computing the sum of differences between the pixels of the colour channels of adjacent fields. While there is some mutual introduction of noise between the colour channels of a single field, the difference between a static and a changing colour channel is still prominent enough. Having found the pattern, we retain only the fields between which all colour channels have changed. For the table above this would mean that we retain the top field of frame 1, the bottom field of frame 2 and the bottom field of frame 3. The result is a progressive video stream with 20 interpolated frames per second, which directly resembles the way the image acquisition system works. The output video therefore follows the following pattern, in which each consecutive frame is a unique RGB colour image:

Frame	1	2	3	4	5	6	...
Red	R_1	R_2	R_3	R_4	R_5	R_6	...
Green	G_1	G_2	G_3	G_4	G_5	G_6	...
Blue	B_1	B_2	B_3	B_4	B_5	B_6	...

Whenever colonoscopic images or videos are used throughout the thesis, they are in this preprocessed state.

The second source of data we have available are measurements of the longitudinal and circular motion of the shaft of the endoscope outside the anus. In a collaborative effort with other researchers in the project group, we have developed a sensor that is able to measure these quantities. The sensor is based on a third-party optical motion sensor, which is, in its current prototype form, spring-mounted in a plastic housing with a channel for insertion of the endoscope. The prototype was designed to be mountable onto the colonoscopy training model M40 by Kyoto Kagaku Co., Ltd., depicted in Figure 3.2. A sensor which can be used in clinical practice is not yet available. However, it is planned and will be developed in the near future.

As with all measurement devices, there is a certain inaccuracy associated with the measurements. On measuring single insertion and withdrawal actions, the device is accurate (relative standard error 0.59 %). However, when accumulated over a colonoscopy procedure, the error can become more significant (relative standard error 8.74 %, compared to twice the extent of the colon - the actual accumulated distance travelled during the course of a procedure, as measured in our experiments, is on average 10.4 times the extent of the colon). Circular motion of the shaft is measured as well and can be very valuable, although with the small circumference of the shaft the error becomes much more significant. Nevertheless can clockwise and counter-clockwise motion be reliably indicated, although the measure is more a qualitative one than a quantitative one.

3.2 Overview of Target Quality Measures

In order to allow for a comparison with existing standards, we choose a subset of the criteria in the JAG DOPS assessment as our target quality measures. To restate our objective from the introduction, we seek to measure quality criteria of the actual colonoscopy procedure, excluding cognitive or diagnostic abilities of the performing endoscopist. In light of this objective, we obtain the following list of target quality measures. The exact wording from the JAG DOPS form follows each measure in parentheses:

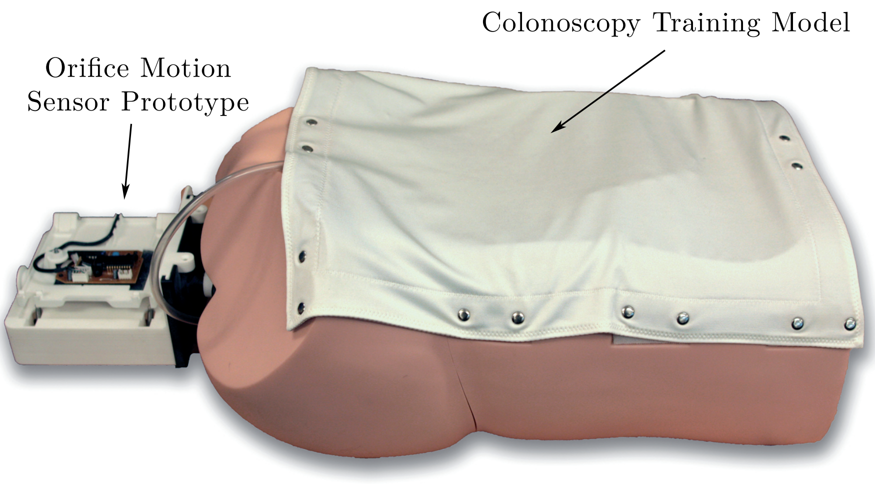


Figure 3.2: Colonoscopy training model M40 by Kyoto Kagaku Co., Ltd., with the motion sensor attached.

- Ability to maintain a clear luminal view. (“Maintains luminal view / inserts in luminal direction.”)
- Endoscope handling. (“Uses torque steering and control knobs appropriately.”)
- Usage of distension, suction and lens washing. (“Uses distension, suction and lens washing appropriately.”)
- Ability to recognise and resolve loops. (“Recognises and logically resolves loop formation.”)
- Usage of position change and abdominal pressure. (“Uses position change and abdominal pressure to aid luminal views.”)
- Appropriateness of procedure time. (“Completes procedure in reasonable time.”)
- Quality of mucosal visualisation. (“Adequate mucosal visualisation.”)

The remaining part of the thesis addresses the problem of automatic measurement of these criteria.

We complement the list of target measures with 3 summary measures for the overall quality of the procedure and its high-level phases:

- Insertion phase performance.
- Withdrawal phase performance.

- Overall procedure performance.

These measures are intended to assess whether there is a difference in performance of the automated assessment system compared to the more specific JAG DOPS criteria.

3.3 Describing Colonoscopy Procedures with Patterns of Image Features and Endoscope Motion

In this section we analyse two texts on colonoscopy screening technique from the major clinical compendium on colonoscopy [139, 63]. Our aim is to identify principal image features with which the stated target measures may be described and quantified. Furthermore, we investigate whether the description of any target measure goes significantly beyond what can be assessed from video and orifice motion data. Such measures are subsequently excluded from further analysis.

Ability to maintain a clear luminal view. This is the major criterion for assessment of the insertion phase. The authors suggest that “the direction of the colonic lumen should be ascertained before pushing in” and establish the rule: “If there is no view, pull back at once.” [63]. These statements suggest that the direction of the lumen is an important characteristic, as well as the reaction of the endoscopist to the event that there is “no view”. No view in this case means a field of view that does not allow an estimation of the camera position for the endoscopist. With regard to our available data, we may be able to detect all the major factors that describe this skill. The *lumen* may be detected in the video frames and it may also be possible to assess the clarity or amount of structure in the field of view from the available video information. Pulling back and insertion of the endoscope is measured directly by our orifice motion sensor.

Endoscope handling. This is marked as a minor criterion in the DOPS assessment form. Torque steering is the preferred method of endoscope steering, and the authors recommend to use “...the lateral angulation control as little as possible.” Torque steering is achieved by first angulating the tip up or down, followed by twisting of the endoscope shaft. Overangulation of the tip should be avoided, as it reduces the ability of the endoscope to slide through the colon. As per this description, torque steering can be measured from video and motion sensor data, in case the angulation of the tip can be determined by camera motion estimation. As stated earlier, the research described in this thesis is part of a broader research project, in which all efforts towards 3D structure and camera motion estimation are on a different branch and not yet developed to a sufficient level. We will therefore analyse the aspect of torque steering using circular motion information only.

Apart from torque steering, the authors advise to "steer slowly and exactly" and "Keep the colonoscope as straight as possible". It is furthermore important to pull back the endoscope "...whenever the view is lost...". For withdrawal, the authors suggest "...constant use of torque...". Again, these aspects may be assessed with camera motion estimation and the motion sensor readings. We see that the loss of a clear view is an important quantity to measure, as it appears again here in connection with endoscope handling.

Usage of insufflation, suction and lens washing. Insufflation stands for the option to inflate the colon with air by pressing a button on the endoscope control section. Another button allows to suction air as well as liquids. Adequate distension of the colon is important, especially for withdrawal, and it is therefore listed as a minor criterion in the DOPS assessment form. If the colon is not sufficiently distended, the walls begin to collapse, possibly covering lesions. On the other hand, an overdistended colon can be very uncomfortable for the patient. Lens washing means to flush the lens in case the view is blocked by liquid or solids on the lens. The authors advise to "insufflate as little as possible" during insertion and "suction air frequently". On withdrawal, it is important to fully suction any liquid faecal pool, as it may cover lesions.

Without direct information about when the buttons are pressed, measuring the usage of insufflation and suction is difficult without a reliable estimate of the 3D structure of the current colon segment, which is beyond the scope of this thesis. A blocked view can be detected by measuring the image clarity. However, we see no possibility to reliably measure whether the low clarity is due to the lens being covered by sticky liquids or solids, without knowing the camera location relative to the colon wall. Due to the strong association with camera position and colon structure, we do not consider the measurement of insufflation, suction and lens washing in this thesis.

Ability to recognise and resolve loops. Loops may form during colonoscopy procedures, especially during the insertion phase. The result is that the force that is applied to the shaft of the endoscope is not transferred to the tip, but instead causes a further widening of the loop. Figure 3.3 illustrates this concept. This can lead to patient discomfort and complications. The ability to recognise and resolve such loops is a major criterion in the DOPS assessment form, as it is a prerequisite for achieving a high completion rate.

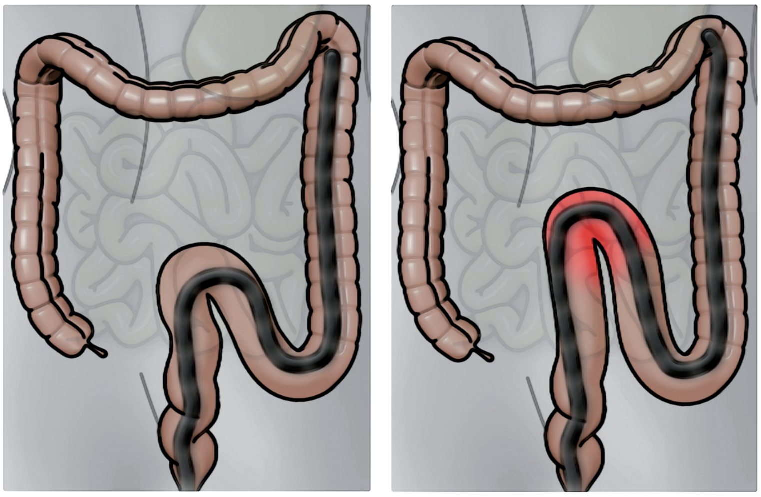


Figure 3.3: Illustration of loop formation in the colon. Forward motion of the shaft of the endoscope at the orifice results in stretching of the colon instead of advancement of the tip of the endoscope.

Loops can be recognised whenever no or little camera motion results from a pushing action. There are certain techniques to resolve loops. For the characteristic *n-loop* that often forms in the sigmoid colon, the authors suggest to "...twist clockwise and withdraw..." in order to resolve it.

While it may be necessary to measure camera motion, the presence of such handling patterns in the motion sensor data may provide a good estimate of the ability to recognise and resolve loops. Hence, we investigate this idea further in Chapter 6.

Usage of position change and abdominal pressure. Changing the patient's position is a strategy to change the configuration of the colon in order to be able to visualise areas which are covered by opaque liquids that can not be suctioned. It may also aid the advancement of the endoscope during insertion. Applying abdominal pressure, according to the authors "...may help modestly during sigmoid intubation, opposing any loop that passes anteriorly,...".

These techniques are neither recognisable from the video data nor from motion sensor information and are therefore not further examined here.

Appropriateness of procedure time. For reasons of patient comfort, "Intubating to the caecum should be as quick as reasonably possible", according to the authors. Trying to hurry intubation is, however, not recommended either. For the withdrawal phase, the authors state that "...optimal detection of lesions requires an adequate amount of time". The optimal duration of withdrawal, however, is an issue that "...no study has adequately addressed...". Appropriateness of procedure time is a minor criterion in the DOPS assessment form.

Measuring insertion and withdrawal time may be achieved by analysing motion sensor readings. We report on this approach in Section 6.2.1.

Quality of mucosal visualisation. Quality of mucosal visualisation is clearly the criterion with the strongest link to overall procedure quality. Perfect mucosal visualisation almost guarantees that no lesion has been missed, unless the endoscopist was distracted or has failed to recognise a lesion as such. Quality of mucosal visualisation is essentially synonymous to withdrawal technique. According to the authors, withdrawal technique involves to "carefully and meticulously examine the proximal sides of the ileocecal valve, all flexures, all haustral folds, and the rectal valves". They advise against a "straight pullback technique" and suggest a "constant use of torque" to examine all space between haustral folds. If the shaft is pulled back too quickly, it is necessary to reinsert and withdraw more carefully, until all mucosal surface has been visualised.

Pushing, pulling and torquing of the endoscope shaft is picked up by the orifice sensor and it can be analysed whether the handling patterns of a good mucosal visualisation are present. With measures of image clarity and known location of the lumen, the ability to automatically assess this criterion is further enhanced. We describe an approach to the measurement of the quality of mucosal visualisation in Chapter 6.

Summary. From this analysis, it appears that the majority of the mentioned criteria can be described by a combination of the following low-level characteristics of colonoscopy procedures:

- Clarity of the endoscopic field of view
- Presence of the lumen in the image
- Postition of the lumen in the image
- Quality of the luminal view
- Motion of the endoscope shaft at the orifice
- Motion of the endoscopic camera

We have stated earlier that we can not address motion of the endoscopic camera in this work. Apart from this, we will present novel approaches to measuring all of the above characteristics in the following chapter.

3.4 Overview of the Complete Approach

In this chapter, we have introduced the type of data we have available, together with necessary preprocessing steps. We have selected a number of target quality measures from the JAG DOPS assessment and additional summary scores and analysed the potential for their automatic measurement. Together with the background information and literature review on the previous chapter, this should provide a broad basis for giving a detailed introduction to our approach towards automatic assessment of individual colonoscopy procedures.

Our approach commences with the measurement of characteristics of single colonoscopic images. Models for these characteristics are learned from an extensive image data set using a universal machine learning framework, which involves automatic feature selection and the training of support vector machines. This is further described in Chapter 4. The next step is to use time and endoscope motion information to derive characteristics of the whole procedure. For this we need a set of procedure videos with synchronised motion sensor readings. In Chapter 5 we describe the design and conduction of an experiment to obtain this data. In the course of this experiment we also collect information on the performing endoscopists and DOPS assessments by two trained experts. Following the development of a large number of procedure characteristics, we select the relevant ones for each of the target measures and train a support vector regression model for each of the target measures. The complete system is shown in Figure 3.4.

In the remainder of this thesis we will describe each of the building blocks of this system in detail, discussing the various stages of the approach shown in the figure from left to right.

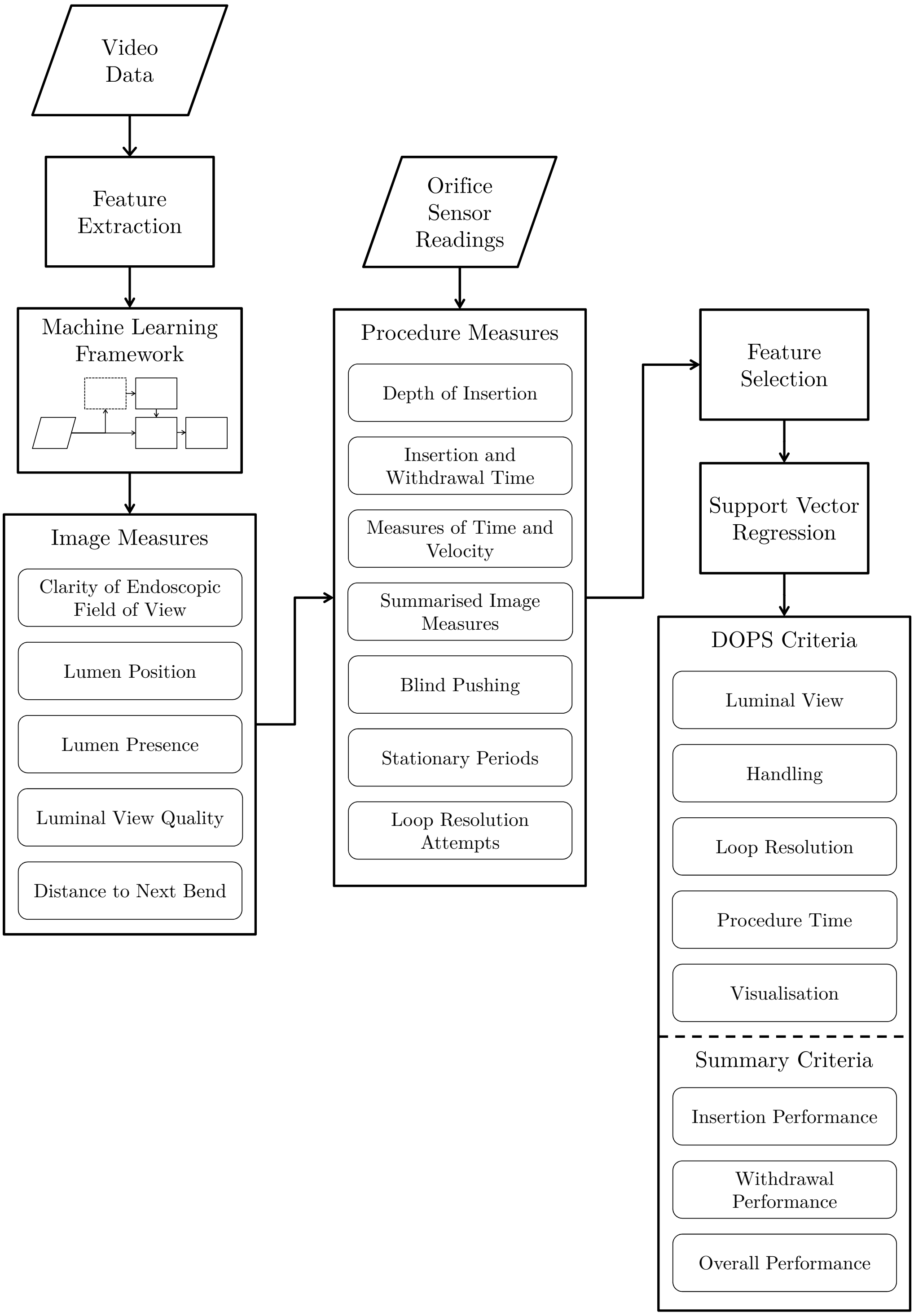


Figure 3.4: Layout of the complete quality assessment system.

Chapter 4

Automatic Measurement of Characteristics of Colonoscopic Images

This chapter comprises the descriptions of methods we developed for measuring characteristics of single colonoscopic images. This is the first set of contributions and forms the basis for the development of higher level measures characterising the whole colonoscopy procedure, which will follow in Chapter 6. We propose a method for the assessment of the clarity of the endoscopic field of view. This topic has previously only been addressed with a binary classification of procedure images into indistinct and informative images. We extend the current state-of-the-art by proposing a clarity measure with multiple grades.

We also introduce measures for different characteristics of luminal views in single images, i.e., luminal view quality, lumen presence, position of the lumen and distance to the closest bend. In measuring these automatically, we achieve a detailed description of colonoscopic images with direct implications to visualisation quality and endoscope handling skills. All image measures are based on models which are trained using a universal machine learning framework involving automatic feature selection and different variants of support vector machines.

Before we go into the measurement of quality related characteristics, we describe a necessary pre-processing step to all the following approaches, which addresses the problem of specular highlights in endoscopic images. Knowledge about these artefacts is crucial for achieving accurate results.

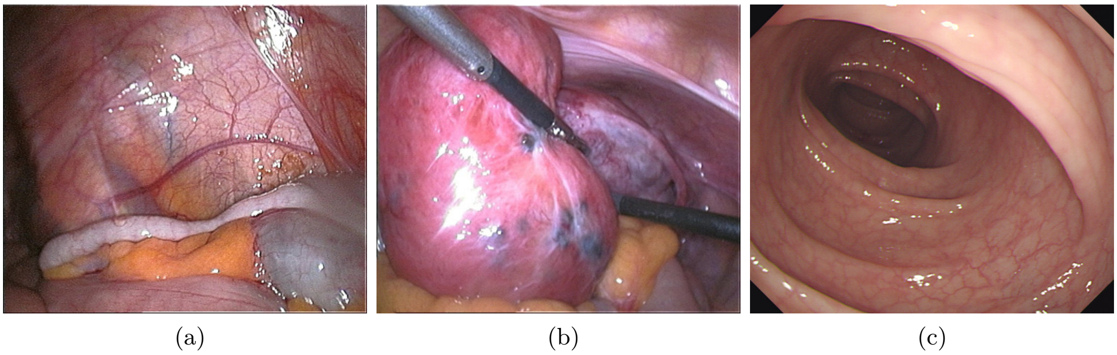


Figure 4.1: Examples of images from minimally invasive medical procedures showing specular highlights. (a) Laparoscopic image of the appendix, (b) Laparoscopic image showing an intervention, (c) Colonoscopic image.

4.1 Detection and Inpainting of Specular Highlights

Images and videos from minimally invasive medical procedures largely show tissues of human organs, such as the mucosa of the gastrointestinal tract in colonoscopy. These surfaces usually have a glossy appearance, showing specular highlights due to specular reflection of the light sources. Figure 4.1 shows example images from different domains with typical specular highlights. These artefacts can negatively affect the perceived image quality [140]. Furthermore, for many visual analysis algorithms, these distinct and bright visual features can become a significant source of error. Since the largest image gradients can usually be found at the edges of specular highlights, they may interfere with all gradient based image analysis algorithms. Similarly, they may affect texture based approaches. On the contrary, specular highlights hold important information about the surface orientation, if the relative locations of the camera and the illumination unit are known. Detecting specular highlights may therefore improve the performance of 3D reconstruction algorithms.

In this section, we propose: (a) a method for the segmentation of specular highlights based on nonlinear filtering and colour image thresholding and (b) an efficient inpainting method that alters the specular regions in a way that eliminates the negative effect on most algorithms and also gives a visually pleasing result.

For many applications, the segmentation will be sufficient, since the determined specular areas can simply be omitted in further computations. For others, it might be necessary or more efficient to inpaint the highlights. We make extensive use of both detection and inpainting of specular highlights throughout the thesis and therefore explain this approach separate from the others without making direct implications to quality measurement.

4.1.1 Detection of Specular Highlights

The proposed segmentation approach comprises two separate modules that make use of two related but different characteristics of specular highlights.

Module 1 The first module uses colour balance adaptive thresholds to determine the parts of specular highlights that show a too high intensity to be part of the non-specular image content. It assumes that the range of colour intensities of the non-specular image content is well within the dynamic range of the image sensor. The automatic exposure correction of endoscope systems is generally reliable in this respect, so the image very rarely shows significant over- or underexposure. In order to maintain compatibility with sequential RGB imaging systems (discussed in Section 3.1), we need to detect specular highlights even if they only occur in one colour channel. While this suggests 3 independent thresholds for each of the 3 colour channels, we set one fixed grey scale threshold T_1 and compute the colour channel thresholds using available image information.

More specifically, the colour channels may have intensity offsets due to colour balancing. At the same time the actual intensity of the specular highlights can be above the point of saturation of all three colour channels. Therefore, we normalise the green and blue colour channels, c_G and c_B, according to the ratios of the 95th percentiles of their intensities to the 95th percentile of the luminance for every image, which we computed as $c_E = 0.2989 \cdot c_R + 0.5870 \cdot c_G + 0.1140 \cdot c_B$ (as defined in [141]) , with c_R being the red colour channel. Using such high percentiles compensates for colour balance issues only if they show in the very high intensity range, which results in a more robust detection for varying lighting and colour balance. The reason why we use the grey scale intensity as a reference instead of the dominating red channel is the fact that intense reddish colours are very common in colonoscopic videos and therefore a red intensity close to saturation occurs not only in connection with specular highlights. We compute the colour balance ratios as follows:

$$r_{GE} = \frac{P_{95}(c_G)}{P_{95}(c_E)} \tag{4.1}$$

and

$$r_{BE} = \frac{P_{95}(c_B)}{P_{95}(c_E)}, \tag{4.2}$$

with $P_{95}(.)$ being the 95th percentile. Using these ratios, any given pixel $\mathbf{x}_0$ is marked as a possible specular highlight if the following condition is met:

$$c_G(\mathbf{x}_0) > r_{GE} \cdot T_1 \quad \vee \quad c_B(\mathbf{x}_0) > r_{BE} \cdot T_1 \quad \vee \quad c_E(\mathbf{x}_0) > T_1. \tag{4.3}$$

Module 2 The second module compares the colour of every given pixel to an estimated non-specular colour at the pixel position. This non-specular colour is estimated from neighbourhood image statistics. This module is aimed at detecting the less intense parts of the specular highlights in the image. Looking at a given pixel, the underlying non-specular surface colour could be estimated as a colour representative of an area surrounding the pixel, if it was known that this area did not contain specular highlights or at least which pixels in the area contain specular highlights. Although we do not know this exactly, we can obtain a good estimate using global image thresholding and paying attention to outliers. Once this estimated non-specular colour is computed, we can determine the class (specular / non-specular) of the current pixel from its dissimilarity to this colour.

The algorithm is initialised by an image thresholding step similar to the one in the first module: Using a slightly lower threshold T_2^{abs}, pixels with high intensity are detected using the condition in (4.3). The pixels meeting this condition are likely to belong to specular highlights, which is one part of the information we need. The actual estimation of the non-specular colour is performed by a modified median filter. Similar non-linear filters have been successfully used for defect detection in images and video (see, e.g., [142, 143]), which is a closely related problem. The median filter was chosen for its robustness in the presence of outliers and its edge preserving character, both of which make it an ideal choice for this task.

We incorporate the information about the location of possible specular highlights into the median filter by filling each detected specular region with the centroid of the colours of the pixels in an area within a fixed distance range from the contour of the region. We isolate this area by exclusive disjunction (XOR) of the masks obtained from two different dilation operations on the mask of possible specular highlight locations. For the dilation we use disk shaped structuring elements with radii of 2 pixels and 4 pixels, respectively.

We then perform median filtering on this modified image. Filling possible specular highlights with a representative colour of their surrounding prevents the filtered image to appear too bright in regions where specular highlights cover a large area. Smaller specular highlights are effectively removed by the median filter when using a relatively large window size w. Figure 4.2 shows an example of the output of the median filter.

Following this, specular highlights are found as positive *colour outliers* by comparing the pixel values in the input and the median filtered image. Among the possible distance measures we found that the maximal ratio of the three colour channel intensities in the original image and the median filtered image produced optimal results. For each pixel location $\mathbf{x}$, this intensity ratio $\epsilon_{\max}$ is computed as

$$\epsilon_{\max}(\mathbf{x}) = \max\left\{\frac{c_R(\mathbf{x})}{c_R^*(\mathbf{x})}, \frac{c_G(\mathbf{x})}{c_G^*(\mathbf{x})}, \frac{c_B(\mathbf{x})}{c_B^*(\mathbf{x})}\right\}, \tag{4.4}$$

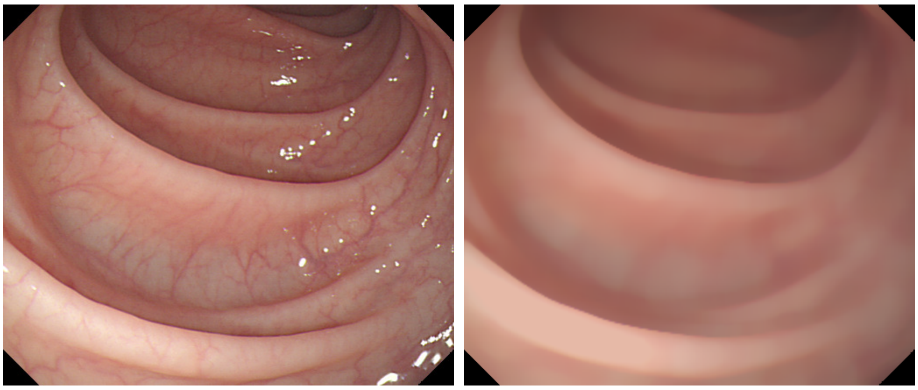


Figure 4.2: Example of a colonoscopic image before and after median filtering.

with $c_R^*(\mathbf{x}), c_G^*(\mathbf{x})$ and $c_B^*(\mathbf{x})$ being the intensities of the red, green and blue colour channel in the median filtered image, respectively. Here again, varying colour balance and contrast can lead to large variations of this characteristic for different images. These variations are compensated using a contrast coefficient τ_i, which is calculated for each of the 3 colour channels for every given image as

$$\tau_i = \left(\frac{\overline{c}_i + s(c_i)}{\overline{c}_i}\right)^{-1}, \; i \in \{R, G, B\}, \tag{4.5}$$

with $\overline{c}_i$ being the sample mean of all pixel intensities in colour channel i and $s(c_i)$ being the sample standard deviation. Using these coefficients, we modify (4.4) to obtain the contrast compensated intensity ratio $\widetilde{\epsilon}_{\max}$ as follows:

$$\widetilde{\epsilon}_{\max}(\mathbf{x}) = \max\left\{\tau_R \frac{c_R(\mathbf{x})}{c_R^*(\mathbf{x})}, \tau_G \frac{c_G(\mathbf{x})}{c_G^*(\mathbf{x})}, \tau_B \frac{c_B(\mathbf{x})}{c_B^*(\mathbf{x})}\right\}. \tag{4.6}$$

Using a threshold T_2^{rel} for this relative measure, the pixel at location $\mathbf{x}$ is then classified as a specular highlight pixel, if

$$\widetilde{\epsilon}_{\max}(\mathbf{x}) > T_2^{\text{rel}}. \tag{4.7}$$

At this point the outputs of the first and second module are joined by logical disjunction (OR) of the resulting masks. The two modules complement each other: The first module uses a global threshold and can therefore only detect the very prominent and bright specular highlights. The less prominent ones are detected by the second module by looking at relative features compared to the underlying surface colour. With a higher dynamic range of the image sensor, the second module alone would lead to good results.

However, since the sensor saturates easily, the relative prominence of specular highlights becomes less intense the brighter a given area of an image is. It is these situations in which the first module still allows detection.

Post-processing During initial tests we noticed that some bright regions in the image are mistaken for specular highlights by the algorithm presented so far. In particular, the mucosal surface in the close vicinity of the camera can appear saturated without showing specular reflection and may therefore be picked up by the detection algorithm. To address this problem, we made use of the property, that the image area surrounding the contour of specular highlights generally shows strong image gradients. Therefore, we compute the mean of the gradient magnitude in a stripe-like area within a fixed distance to the contours of the detected specular regions. Using this information, only those specular regions are retained, whose corresponding contour areas meet the condition

$$\frac{1}{N}\sum_{n=1}^{N} |\mathrm{grad}(E_n)| > T_3 \ \wedge \ N > N_{\mathrm{min}}, \tag{4.8}$$

with $|\mathrm{grad}(E_n)|$ being the grey scale gradient magnitude of the n-th out of N pixels of the contour area corresponding to a given possible specular region. N_{min} is a constant allowing to restrict the computation to larger specular regions, as the problem of non-specular saturation occurs mainly in large uniform areas. The gradient is approximated by vertical and horizontal differences of directly neighbouring pixels. Using this approach, bright, non-specular regions such as the large one on the right in Figure 4.3(a), can be identified as false detections. Figure 4.3 illustrates the idea.

In the presence of strong noise it can happen that single isolated pixels are classified as specular highlights. These are at this stage removed by morphological erosion. The final touch to the algorithm is a slightly stronger dilation of the resulting binary mask, which extends the specular regions more than it would be necessary to compensate for the erosion. This step is motivated by the fact that the transition from specular to non-specular areas is not a step function but spread due to blur induced by factors such as motion or residues on the camera lens. The mask is therefore slightly extended to better cover the spread out regions.

4.1.2 Evaluation of the Segmentation Method

In order to evaluate the proposed algorithm, a large ground truth dataset was created by manually labelling 100 images from 20 different colonoscopy videos. Since negative effects of specular highlights on image analysis algorithms are mostly due to the strong gradients along their contours, the gradient magnitudes were computed using a Sobel operator and

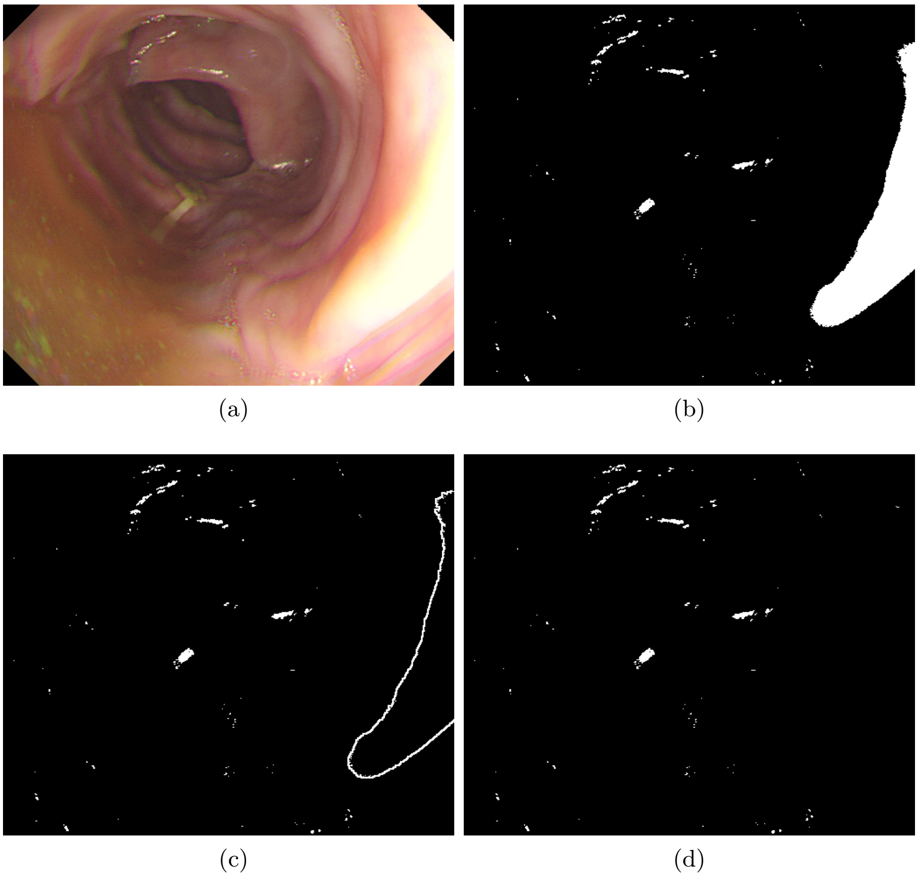


Figure 4.3: Illustration of the area that is used for the gradient test. (a) original image, (b) detected specular highlights, (c) contour areas for the gradient test, (d) resulting specular highlights after the gradient test.

overlayed on the images. This allowed the manual labelling to be very precise on the contours. Great care was taken in including the contours fully in the marked specular regions.

In order to compare the performance of the proposed algorithm with the state of the art, we implemented the approach proposed by Oh, et. al., as described in [54], which was also proposed for detection of specular highlights in endoscopic images. Both methods were assessed by their performance to classify the pixels of a given image into either specular highlight pixels or other pixels.

Using the aforementioned data set, we evaluated both methods using a cross-validation scheme where in each iteration the images of one video were used as the test set and the rest of the images were used as the training set. For each iteration we optimised the parameters of both the method in [54] and the proposed one using a grid search on the training set and tested their performance on the test set. We chose two different

Table 4.1: Performance of the algorithm for equal costs of false positives and false negatives. Compared to the method in [54] with dilation the proposed method achieves a cost reduction of 28.16%. (The abbreviations stand for accuracy, precision, sensitivity and specificity)

Method	**Cost**	Acc. [%]	Prec. [%]	Sens. [%]	Spec. [%]
Oh, et. al.	**8070**	96.83	87.76	37.27	99.25
Oh, et. al. + Dilation	**6473**	97.35	86.66	53.34	99.14
Proposed Method	**4650**	98.33	81.29	75.31	99.28

Table 4.2: Performance of the algorithm for doubled costs of false negatives. Compared to the method in [54] with dilation the proposed method achieves a cost reduction of 31.03%.

Method	**Cost**	Acc. [%]	Prec. [%]	Sens. [%]	Spec. [%]
Oh, et. al.	**15400**	96.70	86.15	39.94	99.01
Oh, et. al. + Dilation	**10271**	97.05	68.85	69.09	98.13
Proposed Method	**7084**	97.90	70.23	83.78	98.51

cost scenarios to measure optimal performance: scenario A assigned equal costs (1 per misclassified pixel) to both missed specular highlights and false positives; scenario B assigned twice the cost to missed specular highlights (2 per missed specular highlight pixel and 1 per false positive).

The results are reported in Tables 4.1 and 4.2 with the resulting cost and the commonly used measures *accuracy*, *precision*, *sensitivity* and *specificity* [144], for the two cost scenarios, averaged over the 20 cross-validation iterations. We report two different variants of the method in [54]. One is the original method as it was reported. The second method is equivalent to the first, followed by a dilation similar to the one in the post-processing step of the proposed method. This was considered appropriate and necessary for a better comparison of the two methods, because in our understanding of the extent of specular highlights, any image gradient increase due to the contours of the specular highlights is to be included during labelling, while the definition in [54] was motivated by a purely visual assessment. The overall improvement resulting from this modification, as it can be seen in Tables 4.1 and 4.2, supports this interpretation.

It can be seen that the proposed method outperforms the one presented in [54] substantially with a cost reduction of 28.16% and 31.03% for cost scenario A and B, respectively. Furthermore, the proposed algorithm was able to process 2.34 frames per second on average on a 2.66 GHz Intel® Core2Quad system - a speed improvement of a factor of 23.8

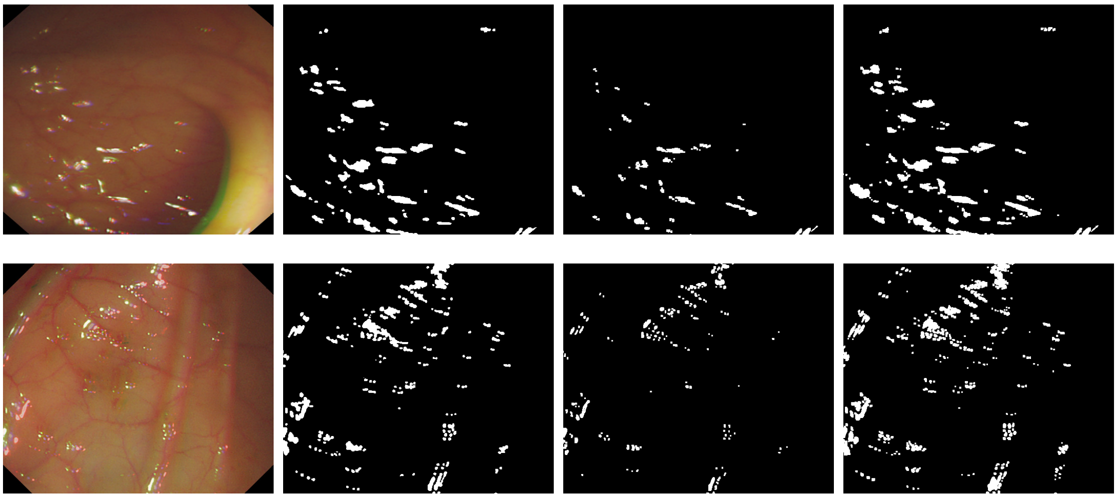


Figure 4.4: Examples illustrating the performance of the specular highlight segmentation algorithm. Original images are shown in the first column. The second column contains the ground truth images, the third column shows the results of the method presented in [54] and in the fourth column the results achieved by the proposed algorithm are depicted.

over the approach presented in [54], which is heavily constrained by its image segmentation algorithm. It took 10.18 seconds on average to process an image. The results are visually depicted for two examples in Figure 4.4.

While the parameters were optimised for each iteration of the cross-validation scheme, they varied only marginally. For images with similar dimensions to the ones used in this study (~528×448 pixels), we recommend to use the following parameters for cost scenario A (cost scenario B): $T_1 = 245\,(240)$, $T_2^{abs} = 210\,(195)$, $T_2^{rel} = 0.95\,(1.00)$, median filter window size $w = 30\,(33)$, $N_{min} = 9460\,(9460)$, $T_3 = 4\,(5)$. The size of the structuring element for the dilation in the post-processing step should be 3 and 5 for cost scenario A and B, respectively.

4.1.3 Inpainting of Specular Highlights

Image inpainting is the process of restoring missing data in still images and usually refers to interpolation of the missing pixels using information of the surrounding neighbourhood. An overview over the commonly used techniques can be found in [145] or, for video data, in [146].

For some applications in automated analysis of endoscopic videos, inpainting will not be necessary. The information about specular highlights can often be used directly (in algorithms exploiting this knowledge), or the specular regions can be excluded from further processing. However, a study by Vogt, et al. [140], suggests that well inpainted endoscopic images are preferred by physicians over images showing specular highlights. Algorithms with the intention of visual enhancement may therefore benefit from a visually pleasing inpainting strategy, as well as algorithms working in the frequency domain. Vogt, et al. [140], also proposed an inpainting method based on temporal information. It can, however, only be only used for a sequence of frames in a video and not for individual images.

Another inpainting method was reported by Cao, et al., in [110]. The authors replaced the pixels inside a sliding rectangular window by the average intensity of the window outline, once the window covered a specular highlight. The approach can not be used universally, as it is matched to the specular highlight segmentation algorithm presented in the same paper.

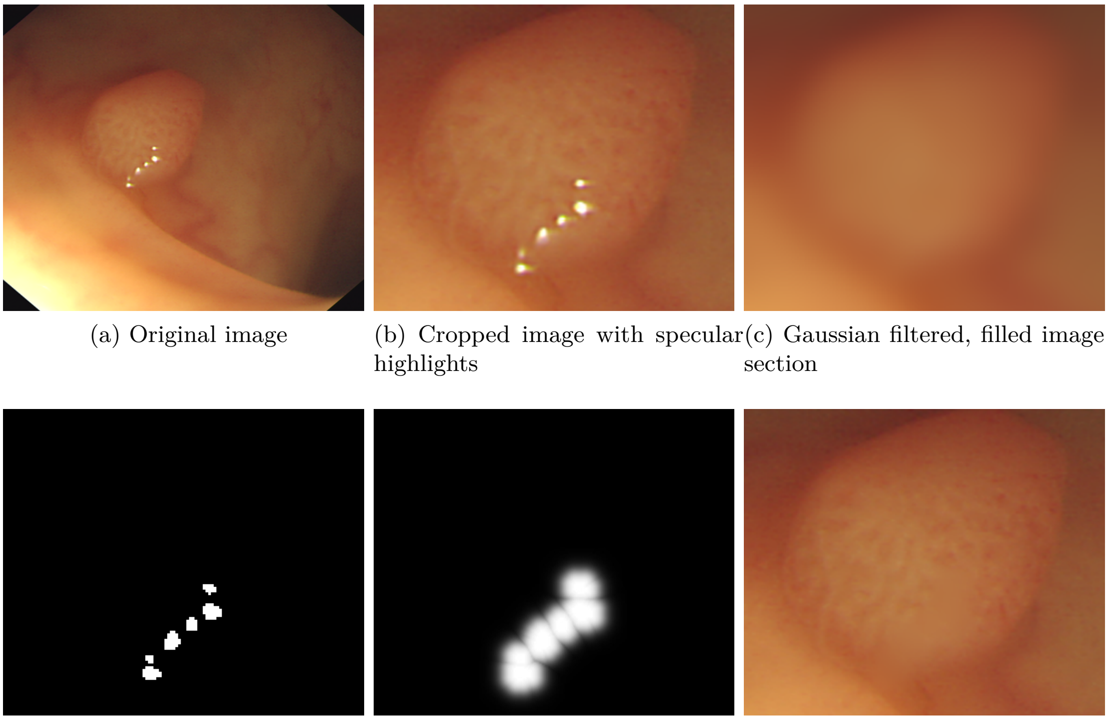
In [54], along with their specular highlight segmentation algorithm, the authors also reported an image inpainting algorithm, where they replaced each detected specular highlight by the average intensity on its contour. A problem with this approach is that the resulting hard transition between the inpainted regions and their surroundings may again lead to strong gradients.

In order to prevent these artefacts, in the proposed algorithm, the inpainting is performed on two levels. We first use the filling technique presented in Section 4.1.1, where we modify the image by replacing all detected specular highlights by the centroid colour of the pixels within a certain distance range of the outline (see above for details). Additionally, we filter this modified image using a Gaussian kernel ($\sigma = 8$), which results in a strongly smoothed image $\mathbf{c}_{\mathrm{sm}}$ free of specular highlights, which is similar to the median filtered image in the segmentation algorithm.

For the second level, the binary mask marking the specular regions in the image is converted to a smooth weighting mask. The smoothing is performed by adding a non-linear decay to the contours of the specular regions. The weights b of the pixels surrounding the specular highlights in the weighting mask are computed depending on their euclidean distance d to the contour of the specular highlight region:

$$b(d) = \left[1 + \exp\left((l_{max} - l_{min})\left(\frac{d}{d_{max}}\right)^{c} + l_{min}\right)\right]^{-1}, d \in [0, d_{max}] . \tag{4.9}$$

This can be interpreted as a logistic decay function in a window from l_{min} to l_{max}, mapped to a distance range from 0 to d_{max}. The constant c can be used to introduce a skew on the decay function. In the examples here, we use the parameters $l_{min} = -5$, $l_{max} = 5$, $d_{max} = 19$ and $c = 0.7$.



(a) Original image

(b) Cropped image with specular highlights

(c) Gaussian filtered, filled image section

(d) Detected specular highlights

(e) Weighting mask

(f) Inpainted image section

Figure 4.5: Stages of the inpainting algorithm.

The resulting integer valued weighting mask $m(\mathbf{x})$ (see Figure 4.5(e) for an example) is used to blend between the original image $\mathbf{c}(\mathbf{x})$ and the smoothed filled image $\mathbf{c}_{\mathrm{sm}}(\mathbf{x})$. The smoothing of the mask results in a gradual transition between $\mathbf{c}(\mathbf{x})$ and $\mathbf{c}_{sm}(\mathbf{x})$. Figure 4.5 illustrates the approach by showing the relevant images and masks.

The inpainted image $\mathbf{c}_{\mathrm{inp}}$ is computed for all pixel locations $\mathbf{x}$ using the following equation:

$$\mathbf{c}_{\mathrm{inp}}(\mathbf{x}) = m(\mathbf{x})\mathbf{c}_{\mathrm{sm}}(\mathbf{x}) + (1 - m(\mathbf{x}))\mathbf{c}(\mathbf{x}), \qquad (4.10)$$

with $m(\mathbf{x}) \in [0, 1]$ for all pixel locations $\mathbf{x}$.

Figure 4.6 shows a number of images before and after inpainting and a comparison to the inpainting method reported in [54]. It can be seen that the proposed inpainting method produces only minor artefacts for small specular highlights. Very large specular regions, however, appear strongly blurred. This is an obvious consequence from the Gaussian smoothing. For more visually pleasing results for large specular areas, it would be necessary to use additional features of the surroundings, such as texture or visible contours. However, such large specular regions are rare in clear colonoscopic images and errors arising from them can therefore usually be neglected. The performance of the combination of the presented segmentation and inpainting algorithms can be seen in an example video which is available online on the following website: `https://www.scss.tcd.ie/~arnoldma/vid_demos/spec.html`.

4.1.4 Discussion

In this section, we have presented methods for segmenting and inpainting specular highlights. We have argued that specular highlights can negatively affect the perceived image quality. Furthermore, they may be a significant source of error, especially for algorithms that make use of the gradient information in an image.

The proposed segmentation approach showed a promising performance in the detailed evaluation. It performed favourably to the approach presented in [54] and avoids any initial image segmentation, thus resulting in significantly shorter computation time (a reduction by a factor of 23.8 for our implementation). Furthermore, in contrast to other approaches, the proposed segmentation method is applicable to the widely used sequential RGB image acquisition systems. The performance of the proposed inpainting approach was demonstrated on a set of images and compared to the inpainting method proposed in [54].

When using inpainting in practice, it is important to keep the users informed that specular highlights are being suppressed and to allow for disablement of this enhancement. For example, while inpainting of specular highlights may help in detecting polyps (both for human observers and algorithms) it could make their categorisation more difficult, as it alters the pit-pattern of the polyp in the vicinity of the specular highlight. Also, as it can be seen in the last row of Figure 4.6, inpainting can have a blurring effect on medical instruments.

4.2 Machine Learning Framework for the Measurement of Image Characteristics

Having discussed video pre-processing, specular highlight detection and inpainting, we now move towards the measurement of the image characteristics we have listed earlier. Those are the clarity of the endoscopic field of view, the presence and position of the lumen in the image, and the quality of the luminal view. We use a common machine learning framework for their measurement, which we present in this section. See Figure 4.7 for a high-level overview of the machine learning framework

The starting point for the learning framework is a set of features obtained from images from M videos of colonoscopy procedures. The actual features were developed through careful analysis of the problem, attempting to encode any image content that possibly

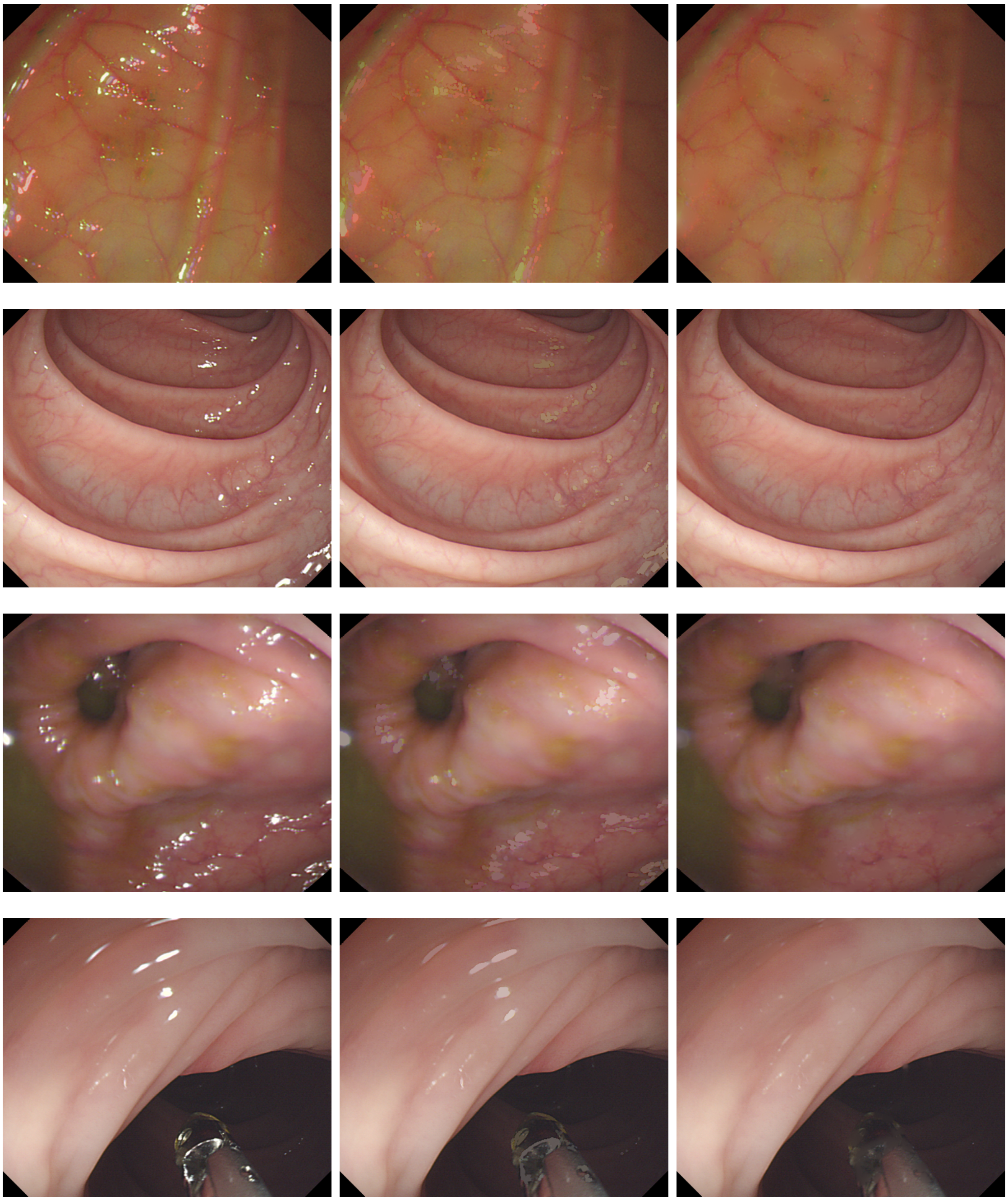


Figure 4.6: Examples illustrating the performance of the inpainting algorithm. Original images are shown in the first column. The second column contains images which were inpainted using the method presented in [54] and the third column shows the results of the proposed method. The segmentation of specular highlights prior to inpainting was performed for both methods using the proposed segmentation algorithm.

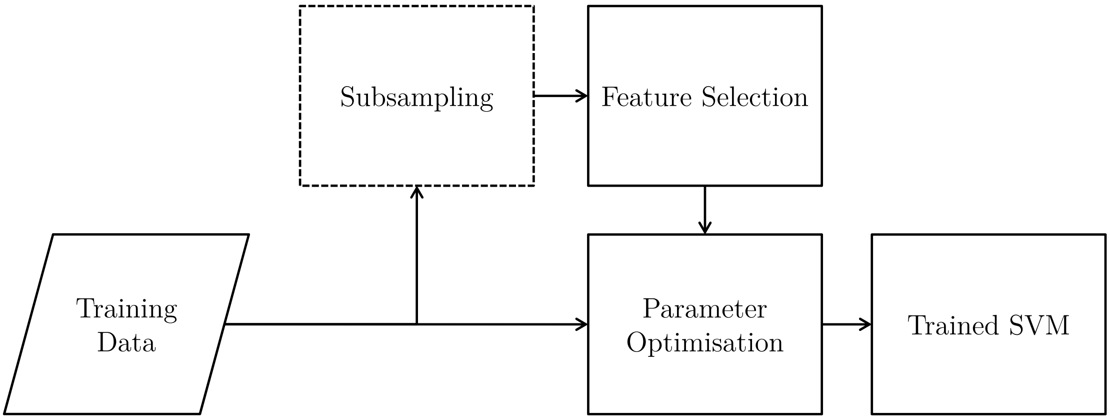


Figure 4.7: High-level overview of the machine learning framework. The use of subsampling depends on the scale of the problem.

relates to the image characteristic in question. The result is a rather large number of features, some of which requiring significant computational resources. Out of these features, many may be redundant and unnecessarily increase the dimensionality of the feature space. We therefore use a feature selection approach to reduce this dimensionality.

4.2.1 Feature Selection

Out of the available methods for feature selection (for an overview, see [147]), we chose a variant of greedy forward feature selection. The method has the advantage that features are not combined or transformed, so that they retain their original meaning, making results more interpretable. Furthermore, removed features do not need to be computed when inference is performed, saving computational resources.

The forward feature selection scheme incorporates parameter optimisation for the inference model in a cross-validation scheme, enhanced by a nested cross-validation (see lower half of Table 4.3) for the best performing features. The nested cross-validation provides a better estimate of the test error to be expected. We compute it in each iteration for the best $K = 3$ features. Depending on the computational resources available, this number can be set higher to improve performance. To ensure validity of the results and avoid overfitting, the cross-validation folds, at any level, split the data in such a way that the validation sets contain data from one video each, and the training sets contain images from all videos except the one of the corresponding validation set. We call this leave-one-video-out cross-validation in the following. A pseudocode description of the feature selection algorithm can be seen in Table 4.3.

Table 4.3: Pseudocode describing the feature selection algorithm used for the assessment of the clarity of the field of view.

```
Start with an empty ordered set of ranked features $\mathcal{F}_r$, the number
  of extracted candidate features $N$ and an ordered set of remaining
  candidate features $\mathcal{F}_c$,
**for** $i = \{1, 2, ..., N\}$ **do**
  **for all** features $f_j \in \mathcal{F}_c$ **do**
    create a set $\mathcal{F}^* = \{f_j, \mathcal{F}_r\}$,
    optimise inference model parameters using leave-one-video-out
      cross-validation and parameter grid search,
    return cross-validation error $\varepsilon_c(j)$ of the optimised inference
      model,
  **end for**
  find the $K$ features $f_k$ yielding the lowest cross-validation errors,
  **for all** features $f_k$ **do**
    create a set $\mathcal{F}^* = \{f_k, \mathcal{F}_r\}$,
    estimate the test error $\varepsilon_t$ of the selected features by nested
      cross-validation (see pseudocode below),
  **end for**
  find the feature $f_{opt}$ yielding the lowest estimated test error,
  remove $f_{opt}$ from $\mathcal{F}_c$ and add it to $\mathcal{F}_r$,
**end for**
find number of features $n_{opt}$, such that $\varepsilon_t(n_{opt}) = \min(\varepsilon_t(n))$.
select the the first $n_{opt}$ features from $\mathcal{F}_r$
```

Nested cross-validation

```
Partition the data into $M$ non-overlapping subsets $\mathcal{V}_m$, each con-
  taining the data of 1 video out of the $M$ videos in the data set,
**for** $k = \{1, 2, ..., M\}$ **do**
  using all $\mathcal{V}_{m \neq k}$, optimise inference model parameters using leave-
    one-video-out cross-validation and parameter grid search,
  train the inference model using the optimised parameters on
    $\bigcup_{m \neq k} \mathcal{V}_m$,
  compute the error $\varepsilon_v(k)$ on video $k$ by applying the trained infer-
    ence model on $\mathcal{V}_k$
**end for**
return the average of $\varepsilon_v$ as the estimated test error $\varepsilon_t$.
```

For large data sets with high numbers of features, this approach becomes computationally prohibitive. Whenever this is the case in the following application of the method, we subsample the training data to a degree that makes computation feasible.

During feature selection we use the same type of inference model as in the final trained system. We use different types of support vector machines (SVM) [148] for the different regression and classification problems. SVMs offer competitive performance and the flexibility to be applicable to a wide variety of supervised learning problems. There exist a number of efficient implementations of the concept. For all SVM computations, we use the LIBSVM support vector machine library [149] in its implementation for MATLAB, with an extension allowing to assign weights to data instances. We use weights to reduce bias effects due to imbalanced training data. We set the weights to be inversely proportional to the number of examples for each grade in the dataset and scale them linearly, such that their average over the data set is 1. However, in the event that subsampling of the training set is performed, it is done randomly, grouped by target value, in order to obtain balanced training sets and avoid the weighting of data instances.

4.2.2 Model Training

Having chosen the features for each of the measures, we train a final optimised SVM in a similar fashion as in the feature ranking algorithm. We use the same type of support vector machine, with parameters optimised through a grid search in a cross-validation scheme. The only difference is that we use a finer grid, because we only need to optimise parameters for a single feature combination. Having found the optimal parameters, the final SVM is trained on all available training data as opposed to the cross-validation subsets that were used up until this point. The model is then applied to test data and its performance is evaluated.

The described machine learning framework will be applied in connection with each of the following image level characteristics. The actual approaches then differ in the features they use, the parametrisation of the framework and the error measures used for optimisation.

4.3 Measuring the Clarity of the Endoscopic Field of View

4.3.1 Clarity in Endoscopic Images

As described in Sect. 2.3.1, the problem of measuring the clarity of the endoscopic field of view has not been previously addressed. Instead, a number of approaches towards detecting indistinct frames have been reported. Detection of indistinct frames can be valuable both as a quality indicator and to reduce the number of images that other algorithms in automated analysis or quality assessment systems need to process. However, it is a matter of definition, which images to regard as indistinct, and this definition may vary depending on the actual application. In the clinical literature, no formal definition of indistinct frames is available. For the use as a quality indicator, one might choose to define as indistinct all images that are blurry to an extent that no clinical diagnosis or characterisation of the mucosal surface is possible. For the technical application as a preprocessing step to an automated analysis system, this definition will have to be altered, as images that are indistinct in a clinical sense may still be valuable for further processing. Figure 4.8 illustrates the problem. In light of the previously described target measures *Ability to maintain a clear luminal view* and *Quality of mucosal visualisation*, and in order to make a useful contribution for all the named applications, we consider it sensible to develop a continuous measure of image clarity. The different definitions of indistinct frames can then be realised by a simple threshold.

4.3.2 Data Set

The model we propose for computation of this measure is learned from a data set that comprises 2627 images from 14 different colonoscopy procedures. We took one sample per second from segments of 120s duration from insertion phases and the same from withdrawal phases. Steering technique differs significantly between in insertion and withdrawal phase, which in turn leads to different viewing angles. To ensure inclusion of the full spectrum of images encountered in colonoscopy procedures, we decided to sample from insertion and withdrawal phases in equal proportion. The images were visualised on a computer screen for manual labelling. We considered image artefacts due to camera defocus and camera motion. In the commonly used sequential RGB imaging systems, camera motion results not only in image blur, but also in colour channel misalignment. The characteristics of these artefacts were used to define the following set of grade descriptors to guide image labelling:

- 4 - Good image clarity, at most minor blurry parts or slight colour misalignment.
- 3 - Acceptable image clarity, either overall blur affecting the inspection quality, or small parts of the image strongly blurred, or medium colour misalignment.
- 2 - Lack of image clarity strongly affects inspection quality, while position of the endoscope can still be well guessed.
- 1 - Completely blurry frame, no inspection possible, no or very vague ability to guess position of the endoscope.

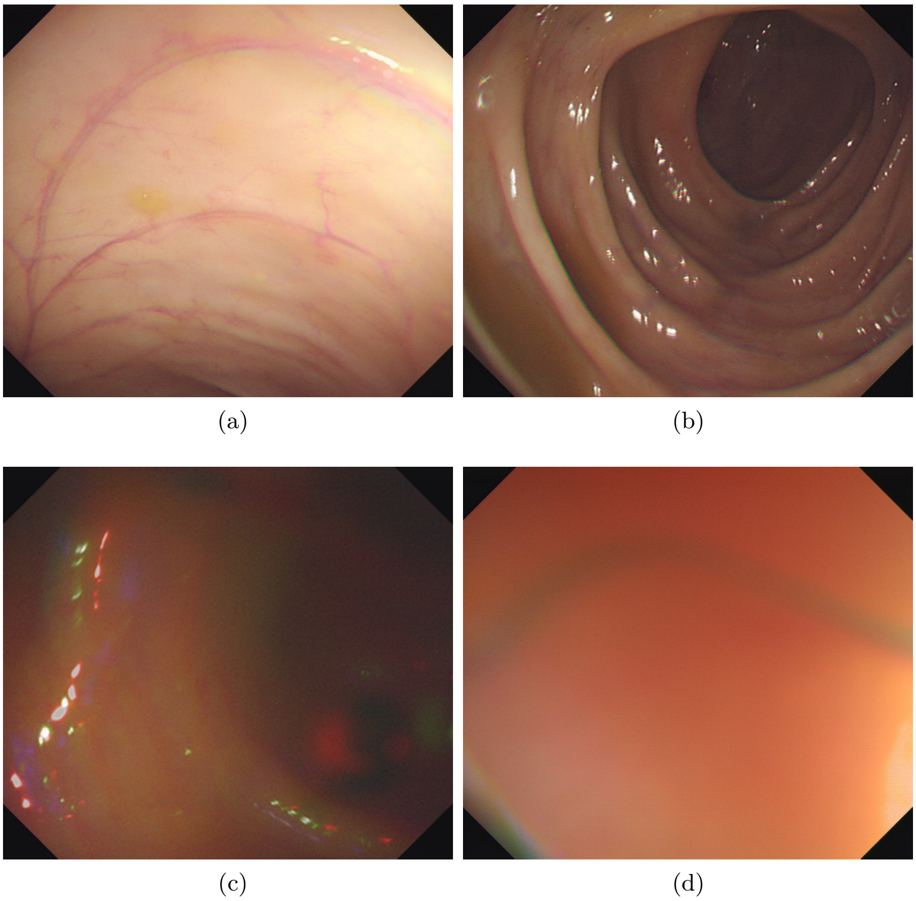


Figure 4.8: Examples of different types of frames. The images in Fig. 4.8(a) and (b) are clearly informative, while the image in Fig. 4.8(d) is clearly indistinct. The blurry image in Fig. 4.8(c), however, may hold important information about the camera motion or the colonic structure, while it does not allow for clinical assessment of the visualised mucosa.

We labelled the images in random order. Consecutive images are often very similar in the data set, which may introduce bias if the image were labelled in their original order. Another reason for random labelling was to spread possible changes of rater bias during the labelling process evenly over the data set.

4.3.3 Method

Previous approaches to detecting indistinct frames, as mentioned in the literature review in Chapter 2, concentrated on the presence of edges in the image [60] or on spatial frequency characteristics [57]. In [53], we have published a method for indistinct frame detection based on the 2D discrete wavelet transform (DWT [150]).

We propose an algorithm that combines the DWT based measure of this earlier publication with a number of additional features. The previously described machine learning framework is applied to obtain a support vector regression model for measuring image clarity. The additional features are the average hue, value and saturation of the image in the HSV colour space, the image contrast, and a novel representation of structure in the image based on intensity histograms.

Feature Computation

The RGB image is first resized to a significantly smaller size, but large enough to be able to distinguish the 4 different clarity grades. Depending on the original image height h and width w, the image is resized such that $\sqrt{h \times w} \approx 256$. It may be justifiable to add another clarity level 5 which would make a higher resolution necessary. However, such very clear images are rare in practice and do not add a significant amount of clinically relevant detail. The fixed image size makes the algorithm invariant to the resolution of the input image.

For the representation of structure in the image, we propose two methods that approach the problem from two different angles. The first representation is the measure we proposed earlier in [53], based on a single level 2D DWT using the Haar wavelet. The DWT results in a set of approximation and detail coefficients. The approximation coefficients represent the low frequency content of the image, while the detail coefficients contain the complementary high frequency information. The proposed DWT feature is obtained by computing the mean of squares of all detail coefficients.

The second representation of image structure we propose is based on histograms. We notate a histogram with k bins as a vector $\mathbf{m}$ with k entries. After conversion to the HSV colour space, we compute histograms $\mathbf{m}_{d,i}$ of each individual row, column and diagonal in the *value* channel of the image. The subscript d is a place-holder for the direction that the histogram is computed for (h: horizontal, v: vertical, $d1, d2$: diagonals 1 and 2), and

$i \in \{1..N_d\}$, with N_d being the number of individual rows, columns or diagonals. This results in 4 ordered sets of histograms, $\mathcal{M}_h, \mathcal{M}_v, \mathcal{M}_{d1}$ and $\mathcal{M}_{d2}$. We use $k = 16$ equally spaced bins, corresponding to a 4 bit encoding of the value information. For all 4 sets, we compute the mean squared differences between adjacent histograms,

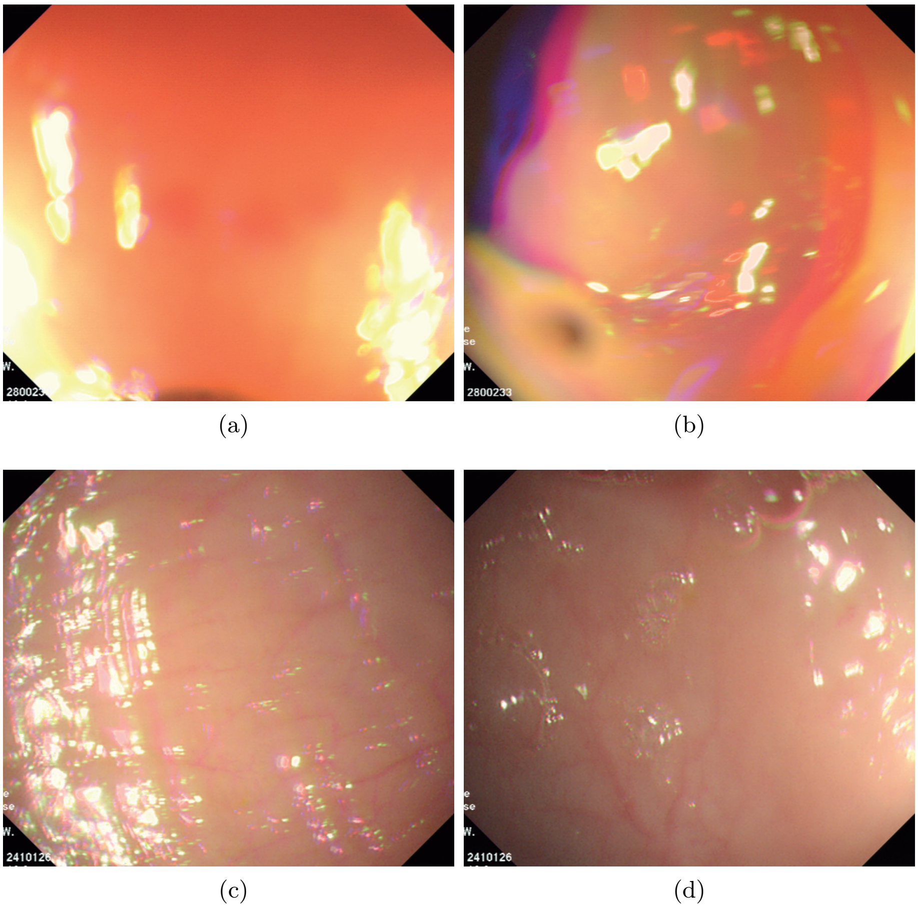
$$\overline{\mathbf{m}}_d = \frac{\sum\limits_{i=1}^{N_d-1} (\mathbf{m}_{d,i+1} - \mathbf{m}_{d,i})^2}{N_d - 1},$$

and, finally, add up all $\overline{\mathbf{m}}_d$ to obtain $\overline{\mathbf{m}} = \sum_d \overline{\mathbf{m}}_d$. Images with little structure and smooth value transitions result in overall low $\overline{\mathbf{m}}$ compared to images with higher clarity and more visible structure. The spread of this information over the k feature bins is what makes this representation valuable, as it describes between which amplitude levels the variations in the image take place. The image is therefore represented by a structure signature which gives the learning algorithm additional cues for discriminating clarity levels.

In the previous section, we have proposed a method for detection and inpainting of specular highlights. We incorporate the method here due to the major impact of specular highlights on gradient based analysis. Specular highlights may appear with sharp edges even in extremely blurry images, yielding higher measurements in our image structure representations than desired. On the other hand, a clear perpendicular view on a smooth part of the mucosal surface may prove to be almost without structure when specular highlights have been removed, as these kinds of images show a large number of small specular highlights due to the uneven mucosal surface. Figure 4.9 shows examples of such images. It is therefore not wise to simply remove specular highlights, as it was done in, e.g., [54], because with their removal the information they hold is removed as well. Ideally, the knowledge about the specular highlights should be incorporated into the algorithm.

In order to do this we compute the structure related features from both the raw images and the images with inpainted specular highlights. Furthermore, we add the number of specular highlight regions and the sum of their areas as features. The optimal set of features is then chosen using the feature ranking algorithm of our proposed machine learning framework.

Apart from image structure and specular highlights, we add a number of image statistics to the set of candidate features, namely the average hue, saturation and value of the image, as well as the image contrast (in the form of the standard deviation of the HSV value channel). These values are included in order to achieve better invariance to lighting conditions and colour balance. They may also help in identifying blurry images, as they often exhibit strongly saturated red colour, hence the term *red-out*, which is often used in clinical literature. Table 4.4 lists all candidate features.



(a) (b)

(c) (d)

Figure 4.9: Examples showing that specular highlights can falsely suggest the presence of structure in the image for gradient based clarity measures when appearing on indistinct frames (a,b). On the other hand, they can be an important cue for informative images showing a very smooth surface (c,d).

Table 4.4: List of features for measuring the clarity of the field of view.

DWT measure (raw image)
DWT measure (non-specular areas)
Mean hue (inpainted image; split into sine and cosine components)
Mean saturation (inpainted image)
Mean value (inpainted image)
Contrast (inpainted image)
Number of specular regions
Overall area of specular regions
Histogram measure $\overline{\mathbf{m}}$ (raw image, 16 features)
Histogram measure $\overline{\mathbf{m}}$ (inpainted image, 16 features)

Feature Selection and Learning of Model Parameters

For application of the machine learning framework presented in Section 4.2, we need to define the type of SVM and the error measure to use for optimisation. Since we have labelled the data using the grades 1 to 4 and are looking for a continuous measure of image clarity, we want to perform regression. For this we choose ν support vector regression with a gaussian radial basis function kernel [148]. This choice leaves us with 3 parameters to optimise: C, ν and the kernel parameter γ. During feature selection, we let $C = 1$ and optimise only ν and γ in order to limit the computational load. We have found in preliminary experiments that this method produces superior results compared to a rougher grid search for all 3 parameters. After feature selection, however, we optimise all 3 parameters before training the final regression model.

The error measure we use for optimisation is motivated by our goal to closely resemble the manually assessed grades of the clarity of the field of view. We therefore use a squared error measure SE that corrects for the discrete character of the rating scale and only counts errors that are still errors after rounding of the predicted value. For a target clarity of $c_t \in \{1, 2, 3, 4\}$ and a predicted clarity of $c_p \in \mathbb{R}$, SE is computed for an example n as

$$SE(n) = \begin{cases} c_p - (c_t + 0.5), & \text{if } c_p > c_t + 0.5, \\ c_p - (c_t - 0.5), & \text{if } c_p \leq c_t - 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{4.11}$$

The rationale behind this error measure is that the discrete rating scale has an underlying continuous range of true values. The rater assigns a discrete value to anything he/she deems to be within the interval $[c_t{-}0.5, c_t{+}0.5[$. We can therefore not know the underlying true value which lies anywhere within this interval, and thus declare every predicted value that results in a correct discrete value after rounding as having zero error. Every predicted value outside the interval is counted as an error with a value relative to the squared distance to the interval.

In order to use this measure for evaluation, it needs to be summarised over all examples it is computed for. A simple approach would be to compute the arithmetic mean over all examples. However, we can achieve better invariance to the distribution of target values, if each group of examples (grouped according to their target values 1,2,3 or 4) is treated separately. The data set can have a strong bias, meaning that one or more groups of examples are over-represented. In case the arithmetic mean is used for optimisation, the performance will be better for the over-represented groups at the expense of the performance on the other groups. To alleviate this tendency, we compute the arithmetic mean of the squared error measure SE for each of the groups separately, resulting in the mean squared error MSE_i for each group i. We then combine the measures of all groups by computing the root mean square (RMS), which further penalises cases in which there is a performance imbalance between groups. The error measure ε for optimisation is therefore computed as

$$\varepsilon = RMS(\langle MSE_i \rangle), i \in \{1, 2, 3, 4\}, \tag{4.12}$$

with

$$RMS(\mathbf{x}) = \sqrt{\frac{1}{N}\left({x_1}^2 + {x_2}^2 + \cdots + {x_N}^2\right)}, \, N = \text{Number of elements in}\, \mathbf{x}. \tag{4.13}$$

4.3.4 Evaluation

To estimate the performance on unseen data, we evaluate the proposed method using the described data set and a cross-validation approach. In each iteration, the data of one video is used for testing, while the rest is used for training the inference model with our machine learning framework. We report the performance using MSE_c defined in 4.11, the Pearson correlation coefficient r and Kendalls's τ statistic. τ is a measure of association regarding only the ordering of the data. Definitions and underlying assumptions of these measures can be found in Appendix A.

The error measure MSE_c, for which the model has been optimised, resulted in a value of 0.0785 in testing. This corresponds to a Pearson correlation of $r = 0.859$, and a Kendall's τ statistic of $\tau = 0.724$. Rounding the predictions to the values in the grading scale, we can directly assess the agreement with the ground truth. The following table shows the percentage of full agreement, complemented by the percentages of predictions having a maximal absolute error of 1,2, and 3, after rounding of the predictions:

Absolute Error	0	≤ 1	≤ 2	≤ 3
% Cases	67.91 %	99.05 %	100 %	100 %

Both Pearson correlation and Kendall's τ suggest a strong association between the predictions and the ground truth. Looking at percent agreement, more than one third of predictions completely agree with the labelled values when being rounded to the grading scale, and 99.05% lie within an error range of ±1.

4.4 Measuring Characteristics of Luminal Views

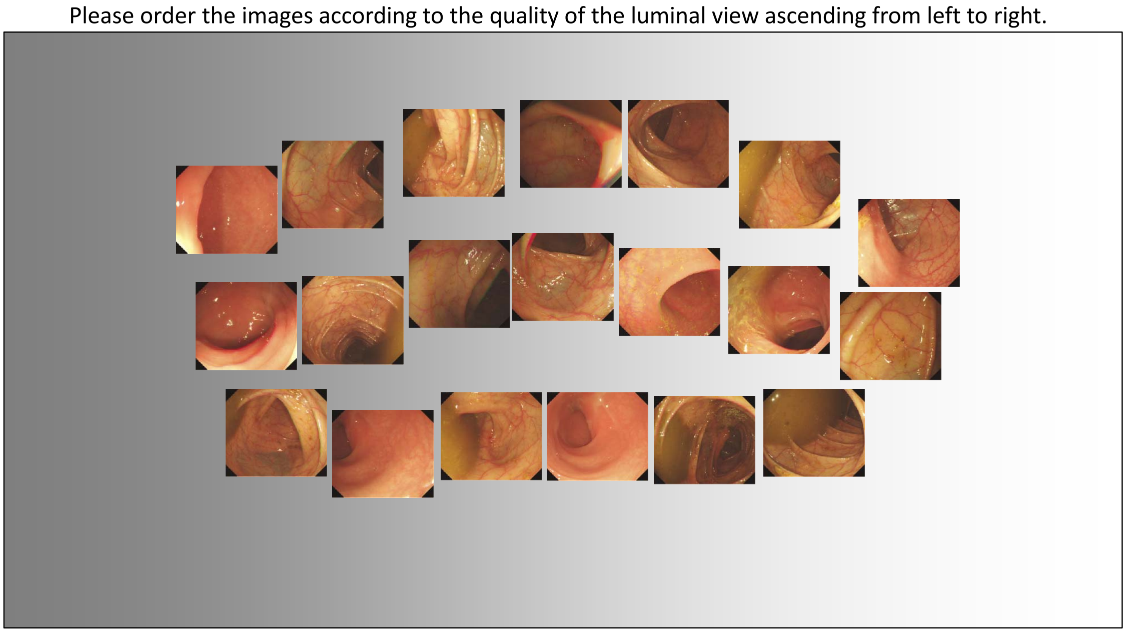
One major quality criterion in the JAG DOPS assessment form is the ability to maintain a luminal view. Unfortunately, there is no official definition of what constitutes a luminal view. In conversations with gastroenterologists, we were told that a luminal view is not just present or absent. There are rather degrees of luminal view quality. The presence of the lumen can be on a range from fully absent to fully present. Other aspects, such as the clarity of the field of view and the proximity to a bend in the colon, also play an important role in determining how good a luminal view is. Previous approaches have not addressed these aspects, which we consider particularly important for achieving clinically significant results.

In this section, we first describe how we arrived at a definition of the quality of luminal views. We then outline the concept of *maximally stable extremal regions*, which we use for segmentation of a number of candidate lumen regions in the image. After this follows a description of the features obtained from these candidate regions and the whole image. We then discuss the parametrisation of the proposed machine learning framework, which we use to train multiple support vector machines for the measurement of various aspects of luminal view quality.

4.4.1 Obtaining a Definition of the Quality of Luminal Views

The definition of the quality of luminal view, which we use in this work, is the result of interviews with two gastroenterology experts. In order to stimulate the thought process we made use of an application in which the experts had to sort a number of example images from colonoscopy videos according to their luminal view quality. Figure 4.10 shows the initial state of the application. The experts had to bring the images in order by dragging and dropping them using the mouse pointer. They were encouraged to explain, during the sorting process, why they made certain decisions for ordering. Figures 4.11 and 4.12 show the final ordering by the two experts. One expert ordered the images from left to right over 3 lines, while the other used the whole height of the background to order images from left to right. Apart from the visual differences, it can be seen that the two experts do not fully agree on the ranking of the images. This may be due to different subjective opinions on what constitutes a good luminal view. Furthermore, images from colonoscopy procedures are so diverse that it is difficult to follow one consistent strategy for ranking.

Once finished, the experts were asked what general characteristics made them rank certain images higher than others. It became apparent that the amount of mucosal surface around the actual lumen largely determines the quality of the luminal view. Images with a centered lumen were rated higher than images with lumen closer to the border of the visual field, because all sides of the current colon segment are visualised. Also, when the viewing distance down the lumen was larger, the images were rated higher than when the



Please order the images according to the quality of the luminal view ascending from left to right.

Figure 4.10: Sorting application in the inital state.

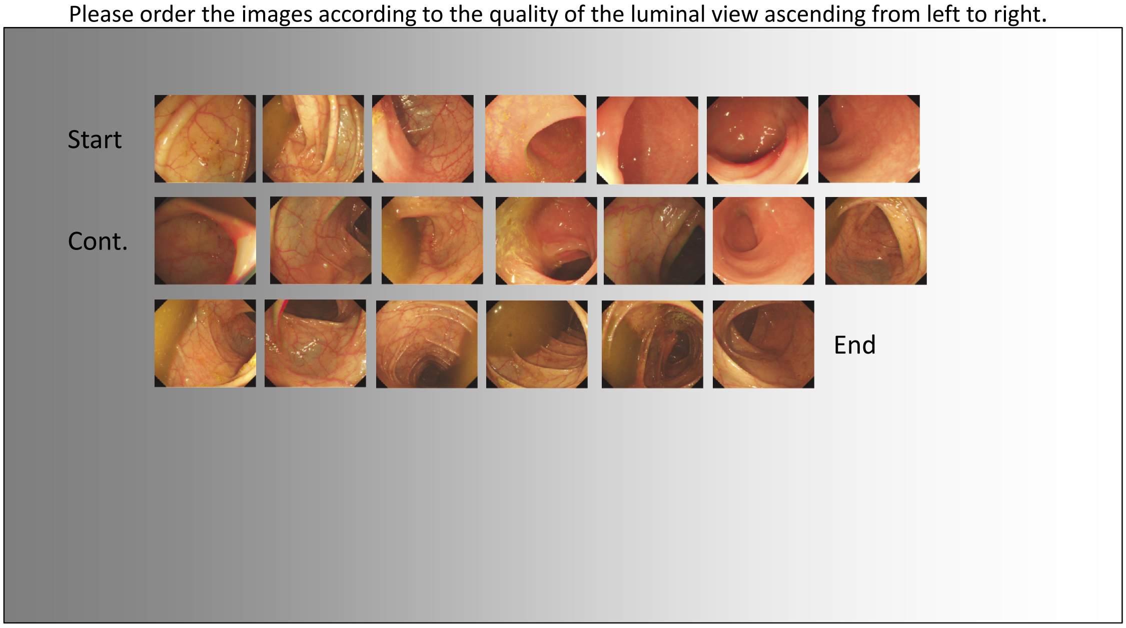
camera was close to the next bend in the colon. Blurry images were not included in the sorting application. However, the experts remarked that for a good luminal view the field of view should be clear.

Given the input from the interview, we developed a rating scale for luminal view quality. The scale ranges from 0 to 4 and the different grades have a description as follows:

4. Good luminal view, endoscope centred or close to centre of tube, with at most slight angular offset. Proximity to a bend does not significantly affect seeing all sides of the surrounding tissue. Clear field of view.

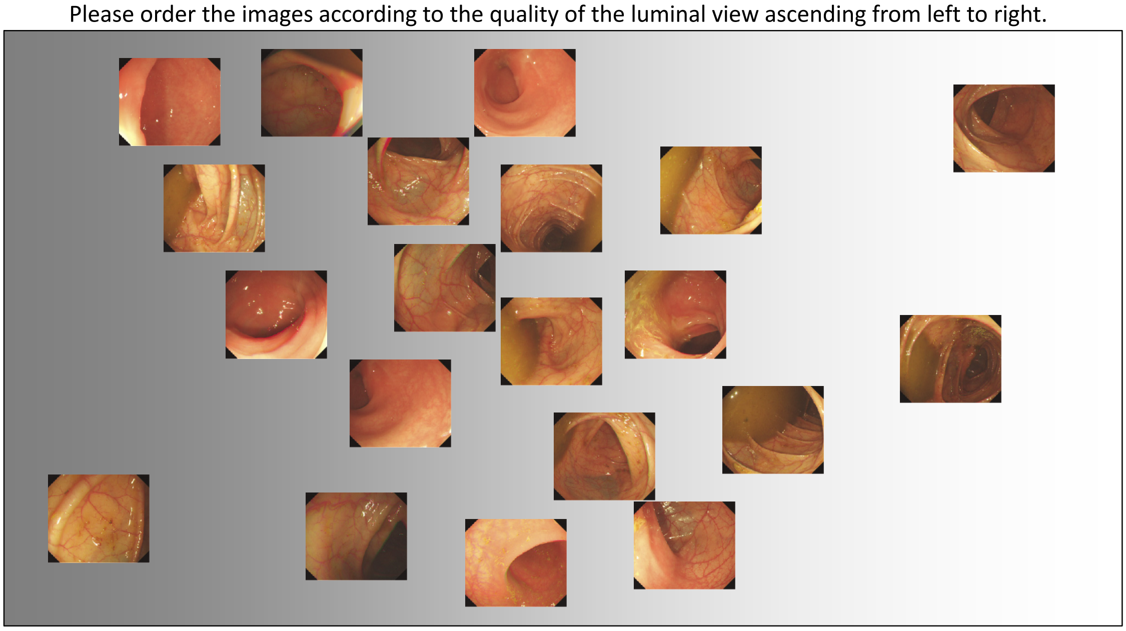
3. Acceptable luminal view that shows the majority of the surrounding tissue, endoscope is noticeably off axis (or centred, on-axis endoscope position, but proximity to a bend or image blur significantly reduces the amount of visualised surrounding tissue).

2. Luminal view with significantly limited visualisation of the surrounding tissue due to considerable off-axis location and/or orientation (or acceptable endoscope position, but proximity to a bend or image blur heavily reduces the amount of visualised surrounding tissue).



Please order the images according to the quality of the luminal view ascending from left to right.

Figure 4.11: Sorting Application: Lumen images as ordered by expert 1. The images were ordered with ascending luminal view quality from left to right over three lines.



Please order the images according to the quality of the luminal view ascending from left to right.

Figure 4.12: Sorting Application: Lumen images as ordered by expert 2. Expert 2 ordered the images, without forming rows, from left to right with ascending luminal view quality.

1. The location and/or orientation of the endoscope is to a degree off-axis that visualisation of the surrounding tissue is strongly limited. However, the farthest point that could be seen from the current camera location (if the orientation were optimal) is still marginally within the field of view (or the next bend is so close, that only a small line segment of the fold on the inside of the bend is visible; or acceptable endoscope position, but image blur heavily affects visibility).

0. The lumen is fully absent, the camera faces the wall almost perpendicularly. The farthest point that could be seen from the current camera location (if the orientation were optimal) is not in the field of view (or indistinct image due to image blur).

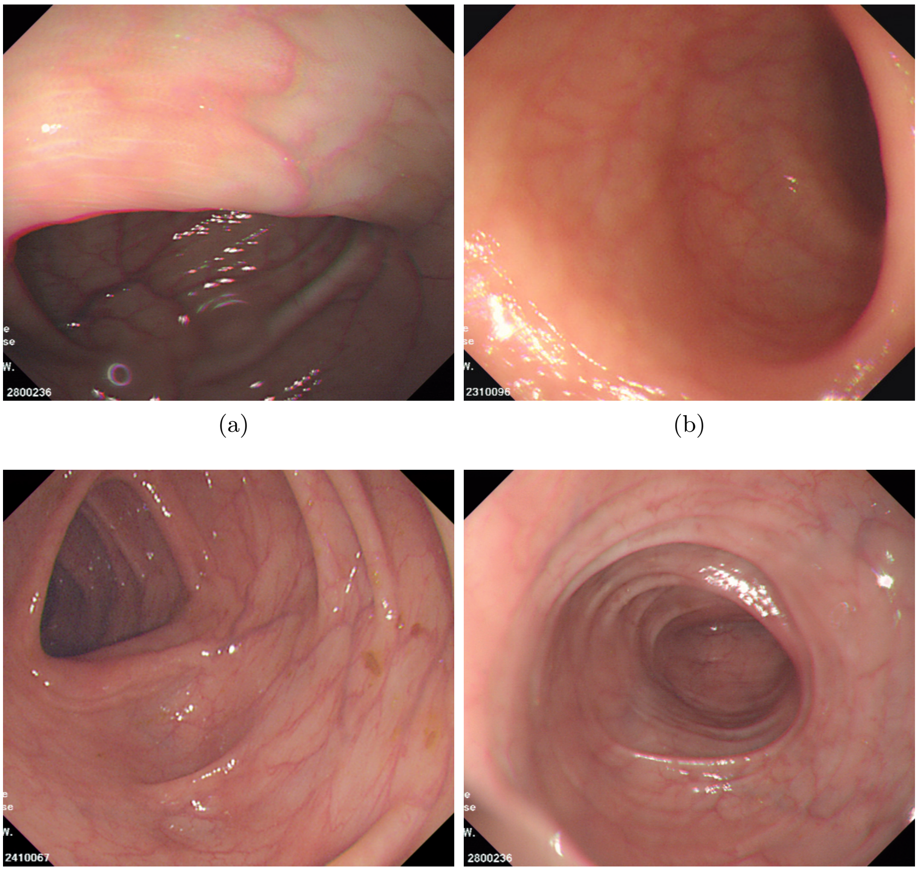
These grade descriptors have been discussed with the experts, who approved them as a representation of their view on luminal view quality. The quality of luminal view is thereby assessed by 3 major criteria: the position of the lumen within the image (and whether it is present at all), the distance to the next bend in the colon, and the clarity of the field of view. All of these are relevant measures on their own for characterising the luminal view more accurately. Therefore, for the labelling of the data set, we used additional rating scales for these measures.

Image clarity was assessed as described earlier in Section 4.3. For the viewing distance, we used a rating scale from 0 to 4, where 0 means the lumen is absent and therefore the camera is facing the wall, and a rating of 3 marks the point where the proximity to a bend starts to affect the visualisation of the surrounding tissue. The position was labelled by marking the centre of the lumen region. Figure 4.13 shows a number of examples with corresponding ratings.

From the amount of detail necessary to describe grades for luminal view quality, and the fact that there is no accurate way of measuring any of the criteria, it is clear that there are limits to the quality of data sets generated with these descriptors. We therefore do not expect very strong agreement between the predictions of our models and the data. Nevertheless, with the expert approval and collaborative exploration of the domain we can assess good face and content validity for the grading, and expect a highly relevant measure for our objective of assessing DOPS criteria in complete procedures.

4.4.2 Maximally Stable Extremal Regions

To provide the necessary background for the description of the algorithm, this section introduces the concept of maximally stable extremal regions. Maximally stable extremal regions (MSER) were introduced by Matas, et. al, in [151] as robust features for wide-



(a)

(b)

(c)

(d)

Figure 4.13: Examples of luminal views with quality grades of a) 1, b) 2, c) 3 and d) 4.

baseline stereo matching. They have since gained widespread attention due to their invariance properties, which make them useful for all kinds of applications that make use of feature matching techniques [152].

Loosely following the definition in [151], extremal regions are sets of spatially adjacent pixels in an image, such that all pixels adjacent to the boundary of that set have an either higher or lower intensity than all pixels in the set. Extremal regions can be found by binarising an image using a threshold, followed by connected component analysis. Varying the threshold t for binarisation of an image yields sequences of nested extremal regions, $\mathcal{R}_1 \subset \mathcal{R}_2 \subset ... \subset \mathcal{R}_N$. Maximally stable extremal regions are defined as extremal regions $\mathcal{R}_i$, for which a threshold change of Δt yields a locally minimal change of the area of the region within the sequence of nested regions.

4.4.3 Overview of the Algorithm

We propose the use of maximally stable extremal regions for segmentation of candidate lumen regions in the image. In contrast to feature matching techniques, it is not the invariance properties that make MSER attractive for this task. When using connected component analysis only for pixels lower than the varying threshold, the method is very efficient in detecting multiple stable dark regions in the image. In preliminary experiments, we noticed that in the vast majority of cases, at least one of those MSER regions provides an adequate segmentation of the lumen. We use intensity, colour and shape features, describing both the regions and the whole image, combined with our measure of image clarity, in order to infer which region most likely represents the actual lumen. Having found the most likely lumen region, the same features help us in measuring 4 different characteristics:

- Position of the lumen: At what location in the image plane is the centre of the lumen?
- Presence of the lumen: Is the lumen within the camera's field of view?
- Distance to the next bend in the colon: To what degree does the distance to the next bend influence the quality of the luminal view?
- Quality of the luminal view: What is the overall quality of the luminal view?

The basic structure of the algorithm is shown in Figure 4.14. The following sections describe the particular implementations of the different building blocks: MSER parametrisation and region filtering, feature extraction, feature selection and learning in further detail.

4.4.4 MSER Parametrisation and Region Pruning

In its standard form, the MSER algorithm has essentially 2 parameters to control its behaviour. The first parameter is the choice whether to perform connected component analysis on the pixels with lower intensity than the moving threshold, or on the ones with greater or equal intensity. Furthermore, a combination of the 2 is possible. The second parameter is the threshold change Δt, for which the stability criterion is evaluated. For our application, the detection of dark regions is desired. We therefore choose to only compute the MSER for intensities below the moving threshold. Regarding Δt, we found a value of $\Delta t = 4$ to yield regions with good lumen representation.

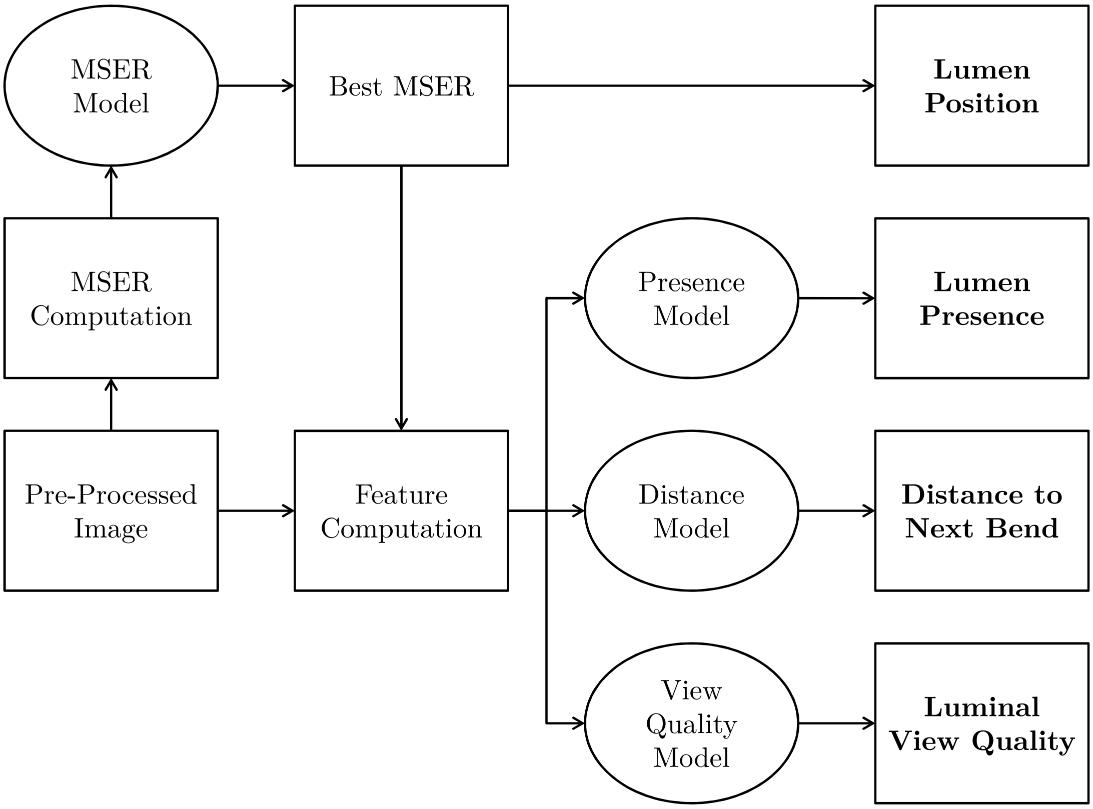


Figure 4.14: Overview of the lumen detection algorithm.

The resulting regions are pruned according to a number of criteria. We set an upper limit of 0.75 on the area of the regions relative to the area of the image, since the lumen area can not cover the whole image plane. We use 0.01 as the lower limit for the area of the regions, meaning that only very small regions get rejected. Good lumen representations may sometimes be obtained by small regions. Hence, we keep this threshold rather low. Furthermore, we want to avoid too similar nested regions and therefore only pick the most stable one in case the relative area variation between a number of nested regions is below a threshold of 0.6.

We use the MSER implementation from the *VLFeat* API [153, 154], which already offers this pruning functionality. The result is an average of 4.6 regions per image, which enter the feature extraction stage of the algorithm.

4.4.5 Feature Extraction and Selection

Having obtained a number of candidate regions, a set of features is extracted from the image as a whole and each of the regions. We incorporate the specular highlight detection method described in Section 4.1.1 at this stage. Specular pixels are not inpainted, but excluded from all feature computations. The features can be divided into 4 different categories.

- **Intensity statistics:** Statistical measures obtained from the image and region intensity. Examples of such features are the mean and standard deviation of the intensity of the regions and the whole image. We use the red channel here, since, for colonoscopy images, it exhibits superior contrast to any of the usual grey scale representations, especially in the darker areas of the image. Intensity is clearly the strongest cue for lumen detection. The lumen region tends to appear darker than the rest of image due to it being the farthest point from the camera and light source.

- **Region shape features:** A set of characteristics of the shape of the region, such as its area or eccentricity. These features are motivated by the fact that shadows behind the illuminated folds in the colon can appear as dark or darker than the lumen. These shadows have a thin and elongated shape and are therefore distinguishable from the usually rounder lumen.

- **Colour statistics:** Basic hue information to highlight differences between region colours and whole image colours. This is mainly to distinguish dark pieces of stool or undigested food from the lumen, as they tend to exhibit a colour that is different from the uniform red-range colour of the colonic mucosa.

- **Image clarity:** The measure of image clarity proposed earlier in Section 4.3.

Table 4.5: List of extracted features for luminal view assessment.

Type	Feature
Clarity	Image clarity measure
Intensity	Mean red channel intensity (Image)
Intensity	Standard deviation red channel intensity (Image)
Intensity	Minimum red channel intensity (Image)
Intensity	Maximum red channel intensity (Image)
Intensity	Median red channel intensity (Image)
Intensity	Lower quartile red channel intensity (Image)
Intensity	Upper quartile red channel intensity (Image)
Intensity	5th percentile red channel intensity (Image)
Intensity	95th precentile red channel intensity (Image)
Intensity	Inter-quartile range red channel intensity (Image)
Colour	Sine of circular mean of the hue (Image)
Colour	Cosine of circular mean of the hue (Image)
Intensity	Mean red channel intensity (Region)
Intensity	Standard deviation red channel intensity (Region)
Intensity	Minimum red channel intensity (Region)
Intensity	Maximum red channel intensity (Region)
Intensity	Median red channel intensity (Region)
Intensity	Lower quartile red channel intensity (Region)
Intensity	Upper quartile red channel intensity (Region)
Intensity	5th percentile red channel intensity (Region)
Intensity	95th precentile red channel intensity (Region)
Intensity	Inter-quartile range red channel intensity (Region)
Colour	Sine of the circular mean of the hue (Region)
Colour	Cosine of the circular mean of the hue (Region)
Shape	Area of the region / Area of the image
Shape	Area of the region when holes are filled / Area of the image
Shape	X position of the centroid of the region / Width of the image
Shape	Y position of the centroid of the region / Height of the image
Shape	Distance of region centroid to image centre / (0.5 * Image diagonal)
Shape	Area within the convex hull of the region / Area of the image
Shape	Eccentricity of the region
Shape	Diameter of the equivalent circle of the region / sqrt(image area)
Shape	Length of the major axis of the equivalent ellipse of the region / sqrt(image area)
Shape	Length of the minor axis of the equivalent ellipse of the region / sqrt(image area)
Shape	Sine of the orientation of the region
Shape	Cosine of the orientation of the region
Shape	Solidity of the region
Shape	Fraction of pixels of the convex hull of the region that are on the image border
Intensity Ratio	Region vs. image: Ratio of mean intensities
Intensity Ratio	Region vs. image: Ratio of standard deviations
Intensity Ratio	Region vs. image: Minimum of region minus minimum of image
Intensity Ratio	Region vs. image: Ratio of maxima
Intensity Ratio	Region vs. image: Ratio of medians
Intensity Ratio	Region vs. image: Ratio of lower quartiles
Intensity Ratio	Region vs. image: Ratio of upper quartiles
Intensity Ratio	Region vs. image: Ratio of inter-quartile ranges
Colour	Region vs. image: Absolute difference of mean hue angles

Table 4.5 shows the complete list of features. The list, as it is, contains many supposedly redundant features. Which features are useful for the luminal view assessment is decided by the feature selection in the machine learning framework described earlier. Feature selection is performed for MSER region assessment and each of the presence, distance and luminal view quality measures individually. Hence, the machine learning framework is parametrised differently for the different measures, varying in the choice of error measure and the type of support vector machine used.

Ranking of MSER Regions. Assessing the MSER regions' degree of representing the lumen region is the basis for all the other measures. Correct ordering of the MSER regions is here more important than approximating their true value. Therefore, we use an error measure based on Kendall's τ statistic (see Appendix A for the definition). Kendall's τ, in its basic form, has the problem that in data sets in which one value is heavily over-represented, the mutual ordering of the other values has only a marginal influence on the statistic. Our data set consists of targets on an integer valued scale from 0 to 4, in which regions with a grade of 0 are over-represented at an approximate ratio of 5:1 compared to the regions with grades from 1 to 4. We therefore compute τ for 5 different subsets $\mathcal{D}_i^*$ of the validation data, with each subset being the full set $\mathcal{D}$ without the examples of one of the 5 different grades. The error measure ε is then obtained subtracting the harmonic mean of the values $\boldsymbol{\tau} = \{\tau_i\}$, $i \in \{0, 1, 2, 3, 4\}$ of the subsets from 1, i.e.,

$$\varepsilon = 1 - \text{harmmean}(\boldsymbol{\tau}), \tag{4.14}$$

with

$$\text{harmmean}(\mathbf{x}) = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}, \; N \in \mathbb{N}, x_i > 0 \,.$$

While we are maximising harmmean($\boldsymbol{\tau}$), it may happen that an element of $\boldsymbol{\tau}$ is zero or negative, for which the harmonic mean is undefined. In this case we replace the harmonic mean by the arithmetic mean and penalise the result by multiplying by 0.1. This case only occurs in the first iteration of the feature selection algorithm and only with the weakest features.

As our inference model, we use ν support vector regression with gaussian radial basis function kernel. We limit predictions to the range of the grading scale, i.e., predictions smaller than 0 or greater than 4 are replaced by 0 and 4, respectively. As it was done earlier for our clarity measure, we keep the cost parameter C fixed at 1 during feature selection and optimise only ν and the kernel parameter γ. Having selected the features, all 3 parameters are optimised before training the final model.

Lumen Presence. Assessing the presence of the lumen is equivalent to discriminating between images with a luminal view quality of 0 (lumen is not present) and all other images (lumen is present). Since this is a binary decision, we use a C-SVM classifier in our machine learning framework, again with a gaussian radial basis function kernel. For the classifier we have to decide for a desired operating point and choose an appropriate error measure to train the classifier accordingly.

In our application, we consider the false positive rate (FPR) to be equally important to the false negative rate (FNR). We could therefore choose to minimise the average of the two, which would be equivalent to maximising the balanced accuracy (BA, defined as $BA = 0.5 \cdot (\text{Sensitivity}+\text{Specificity}) = 1 - 0.5 \cdot (\text{FPR}+\text{FNR})$). However, this average could be optimal despite a large absolute difference between FPR and FNR. We encourage this difference to be lower by computing the root mean square (RMS ((4.13))) instead of the arithmetic mean:

$$\varepsilon = RMS(\langle \text{FPR,FNR} \rangle). \tag{4.15}$$

Distance to the Next Bend. The ground truth data for assessment of the distance to the next bend was labelled on an integer scale from 0 to 4. We use again a ν-SVM regression model and limit the real-valued predictions to the same 0 to 4 range. We want to predict values as close to the ground truth as possible and therefore use again the mean squared error measure introduced in Section 4.3 (equation (4.12)) for optimisation and performance assessment.

Luminal View Quality. The feature ranking for assessment of the quality of the luminal view is treated exactly as for the assessment of the viewing distance, described in the previous paragraph. We use a ν-SVM regression model with the same discrete scale error measure.

4.4.6 Model Training and Application

Having chosen the features for each of the measures, parameter optimisation and model training is performed according to the description in Section 4.2. When applied, the predictions of the regression models are limited to the range of the labelling scales, i.e., predictions greater than 4 are replaced by values of 4 and negative predictions are replaced by zeros.

4.4.7 Evaluation

The data set is the same as the one used for the evaluation of our proposed image clarity measure (2627 images from 14 colonoscopy procedures), containing one sample per second from segments of 120 s duration from insertion phases and the same from withdrawal phases. Similar as for the image clarity computation, sampling equally from insertion and withdrawal phases ensures that all types of luminal views are represented to a sufficient degree in the data set.

We performed the labelling according to the rating scales discussed above. Images were visualised on a computer screen and assessed according to viewing distance and overall luminal view quality. The subjectively perceived lumen position was marked using the mouse pointer. Subsequently, the detected MSER regions were displayed one by one, overlayed on the images, and rated according to their degree of representing the lumen region. We used the following grade descriptors for this assessment:

4. Good to ideal matching with at most a number of holes in the MSER region.

3. Good matching with slightly more or slightly less than the actual lumen covered, or very good matching of a superset of the lumen area that can be considered an alternative, bigger lumen area.

2. MSER region contains the lumen region but also a considerable part of other areas. However, the major part of the MSER region is in the lumen area, or partial matching of a superset of the lumen area that can be considered an alternative, bigger lumen area.

1. The lumen region is covered by the MSER region but about half of the MSER regions area is not part of the lumen area.

0. The MSER region does not cover the lumen area at all or the lumen area is only a minor part of the MSER region.

The result are 12190 labelled regions.

Training and test sets were created using the leave-one-video-out cross-validation approach we described in Section 4.2. All feature selection and parameter optimisation was done according to the machine learning framework we introduced earlier. Preferences in feature selection and overall results are listed for each measure in the following paragraphs.

Ranking of MSER Regions. For the ranking of the MSER regions the feature selection algorithm selected on average 18.9 features. Especially region intensity and intensity ratios between regions and the whole image were found to be most discriminative, together with a number of region shape features. The method achieved a value of 0.373 for our chosen error measure ε (4.4.5). This corresponds to a Kendall's τ coefficient of 0.668 and a Pearson correlation of $r = 0.845$. For definitions of these measures of association see Appendix A. In the manual assessment, 900 images in the data set were marked as containing the lumen. In these images, the highest ranked lumen region, as determined by the proposed method, was identical to the ground truth in 787 cases.

Presence of the Lumen. For the only binary classification task, assessing whether the lumen is present or not, the feature selection algorithm selected on average 15.9 features. Most used features were absolute intensity statistics of the best region and the whole image, together with region shape features, image clarity and region hue. The chosen error measure ε (4.15) amounted to 0.147. This corresponds to a balanced accuracy of 85.5 % and an area under the ROC curve of 0.934. False positive and false negative rates were 17.4 % and 11.6 %, respectively. While the performance may seem a bit low, we have to keep in mind that the lumen is not an object with defined boundaries. Hence, it is, for the human rater, not a straightforward decision whether the lumen is present in the field of view or not.

Viewing Distance Towards the Lumen. For the assessment of the viewing distance the feature selection algorithm chose 9.9 features on average, containing intensity ratios, the centroid of the best region and image clarity among the most prominent ones. Our optimisation performance measure, the RMS of the squared error measure defined in (4.12), amounted to 0.449. This corresponds to a Kendall's τ coefficient of 0.581 and a Pearson correlation of $r = 0.751$. This can be considered a moderate association between the predictions and the ground truth. Tabulating the percentage of predictions within different error bounds further illustrates that the deviation from the ground truth is indeed low for a large number of cases and large deviations are very rare:

Absolute Error	0	≤ 1	≤ 2	≤ 3
% Cases	49.75 %	83.40 %	98.24 %	100 %

Luminal View Quality. The quality of the luminal view was assessed using a set of 13.5 features on average, with image and region intensity features, region shape features, and region centroid being the most prominent ones. The error measure used for optimisation was, again, the RMS mean square error measure defined in (4.12). The regression model achieved an error measure of 0.297, corresponding to a Kendall's τ coefficient of 0.598 and a Pearson correlation of $r = 0.781$. Tabulating the percentage of predictions within different error bounds shows that deviations from the ground truth are small and the automatic measure achieves good agreement with the human rater.

Absolute Error	0	≤ 1	≤ 2	≤ 3	≤ 4
% Cases	57.48 %	91.32 %	98.74 %	99.92 %	100.00 %

Lumen Position. The lumen position is computed as the centroid of the best ranked region. The ideal case for the automated system would be to detect the best regions identical to the ground truth. However, the ground truth lumen positions were marked, independently of any regions, as what the rater considers the centre of the lumen. Therefore, even in the ideal case, there can be no perfect agreement between the detected lumen position and the ground truth. We report the results as the median euclidean distance to the manually marked lumen position, normalised such that a detected lumen position with maximal possible distance to the ground truth position has a distance of 1. Hence, we divide the euclidean distance by the length of the image diagonal. This is for the results to be comparable between images of different size. Our method achieves a median distance of 0.197. For comparison, the ideal choice of regions would yield a median distance of 0.092, while a random choice of the lumen position (X and Y coordinates drawn from independent uniform distributions limited to the image dimensions) would result in a distance of 0.341 (mean of 10000 trials for each image).

4.4.8 Discussion

Characteristics of luminal views are difficult to assess, both for humans and automated systems. This can be seen from the way the two experts have ranked the examples, which we have shown in the introduction to this section. Nevertheless are these characteristics important for assessing a number of quality related aspects of colonoscopy procedures. For example is a good luminal view a prerequisite for a safe insertion phase, and during withdrawal, the lumen is an important feature to identify which side of the colon wall the endoscope is pointing towards.

In an effort to reduce subjectivity and increase consistency, we have introduced rating scales with detailed grade descriptors for the different characteristics of luminal views, supported by a study based on an image ranking task and interviews with two domain experts. Models for the characteristics were trained using the proposed universal machine learning framework and an extensive data set labelled according to the rating scales.

The results show that the proposed methods achieve a promising degree of association with the labelled values. Future research may seek to improve the results by looking at the way the measures evolve over time and possibly apply smoothness constraints (time domain filters).

Chapter 5

Data Collection and Analysis of Expert Assessments

5.1 Introduction

The previous chapter introduced methods for the measurement of various characteristics of colonoscopic images. In order to measure relevant characteristics of complete colonoscopy procedures, the image measures need to be combined with time measurements and readings of the orifice motion sensor. We therefore need to obtain this data for complete procedures simultaneously with the video data. In this chapter, we describe the design of a task driven experimental setup, which allows this simultaneous data recording. In this experiment, endoscopists with various grades of experience perform screening procedures on a colonoscopy training model while being recorded. In addition to video and sensor data, we collect information on the experience of the endoscopists. Due to our collaboration with two endoscopy suites in two different hospitals, we have a reasonably sized pool of endoscopists available. Furthermore, we have access to two assessors who have completed a *train the trainers* course offered by the JAG and are therefore qualified for DOPS assessment. All procedures are assessed according to our selected JAG DOPS criteria and summary scores by the two experts. These assessments are later used as regression targets for the mapping of procedure characteristics to DOPS criteria.

In the following we describe the experimental setup and provide details on the way the screening task and assessment were structured. The collected procedure data is then analysed and interpreted before looking at the inter-rater reliability between the two expert assessors.

Ethics approval for the experiment was obtained from Trinity College Dublin and the contributing hospitals.

5.2 Experimental Setup

5.2.1 Screening Procedures

We chose to record the screening procedures for later offline assessment by the two experts. This allowed us to anonymise the participating endoscopists, so that the assessors had no knowledge of the status or experience of the participants. The downside of offline assessment is that it is a deviation from the original DOPS protocol, where assessors need to be present during the procedure, with possible impacts on the precision of the test. However, since one of the assessors rated endoscopists from the endoscopy unit they work in, we considered the anonymisation to be more important.

Technical Setup. All procedures were performed on a colonoscopy simulator, mainly due to the motion sensor being in its prototype state and optimised for use with the colonoscopy training model M40 by Kyoto Kagaku Co., Ltd.. This model consists of an emulated large intestine, made from soft resin, in a firm housing in the shape of a human torso from the lower ribs to the beginning of the thighs. The layout of the lower intestine model can be adjusted to various difficulty levels using interchangeable layout cards and elastic bands. During the procedure, the housing is covered with elastic material that allows to administer abdominal pressure. Turning the model on its sides is also possible and encouraged in difficult situations. The model can be made airtight during the procedure so that insufflation can be used as in a real procedure. Unfortunately, the procedures could not be performed in fully equipped endoscopy theatres, which prevented the use of suction. Nevertheless, we received much positive feedback on the realism of the training model. The lack of suction was also known to the assessors. We therefore assume only a minor impact on the precision of the assessment.

The colonoscopy training model allowed us to keep a fixed layout of the colon and therefore also a fixed procedure difficulty for all participants. In this case, the DOPS assessment differentiates between the participating endoscopists according to their screening technique, eliminating the impact of anatomical differences between patients and varying bowel preparation quality. We chose a relatively easy layout of the colon model to obtain a large number of completed procedures

In our experimental setup, we used an Olympus standard definition endoscopy system with sequential RGB image acquisition. The video from the endoscope camera was recorded with a PC-based video grabbing system and compressed using a lossless RGB video codec. The signal from the motion sensor was recorded on a second PC. Furthermore, for offline assessment, it is necessary to provide a view of the scene to the assessors, showing the endoscopist's handling of the endoscope and the model. We therefore recorded the

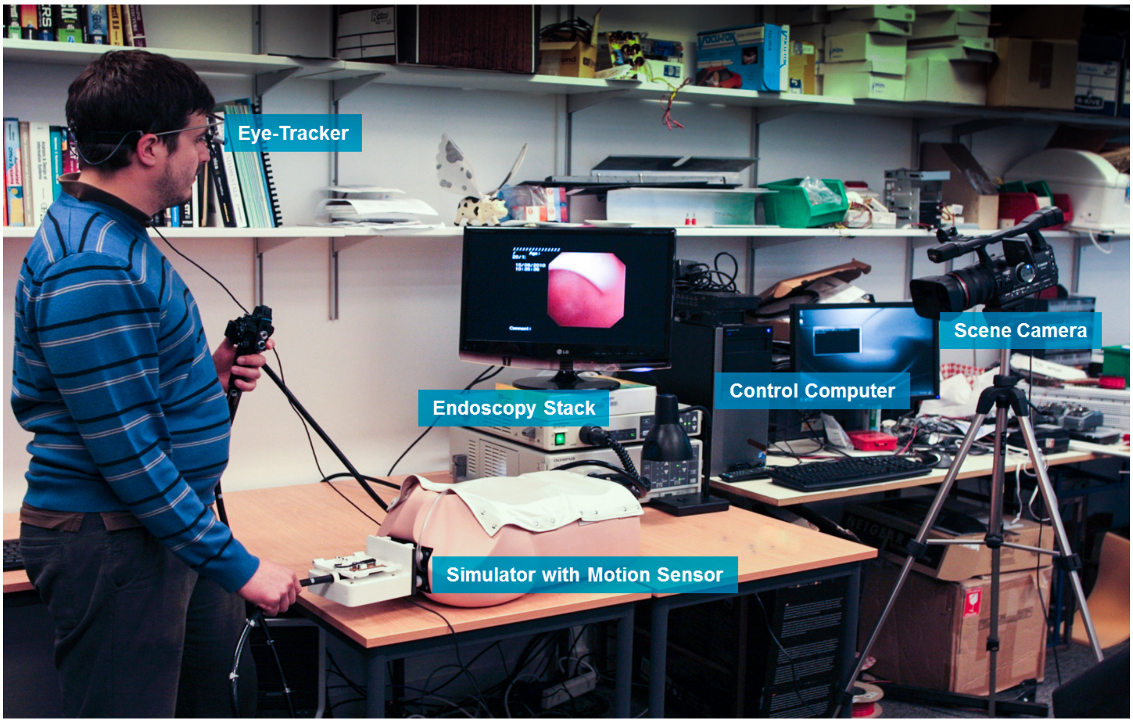


Figure 5.1: Technical setup for the procedure side of the data collection.

scene with an additional camera. We kept the heads of the participants out of the field of view of the scene camera and asked them to wear neutral gowns so that they remain anonymous to the assessors. For additional studies within our research project, the endoscopists also wore an eye-tracking device, again involving two cameras and another video grabbing PC.

For the analysis it was important to maintain accurate synchronisation between the recording devices. We synchronised all computer clocks and used a visual signal as a synchronisation event for time alignment of the external camera. Since the internal computer clocks are of low quality, we measured the drift between the different clocks and subsequently adjusted for it. This drift was as large as up to 4 frames per hour of recording, given the frame rate of 25 frames per second.

The resulting, synchronised data consists of the following:

- Video data from the endoscope.
- Video data from the scene camera.
- Longitudinal motion of the endoscope shaft at the orifice.
- Tangential motion due to rotation of the endoscope shaft at the orifice.

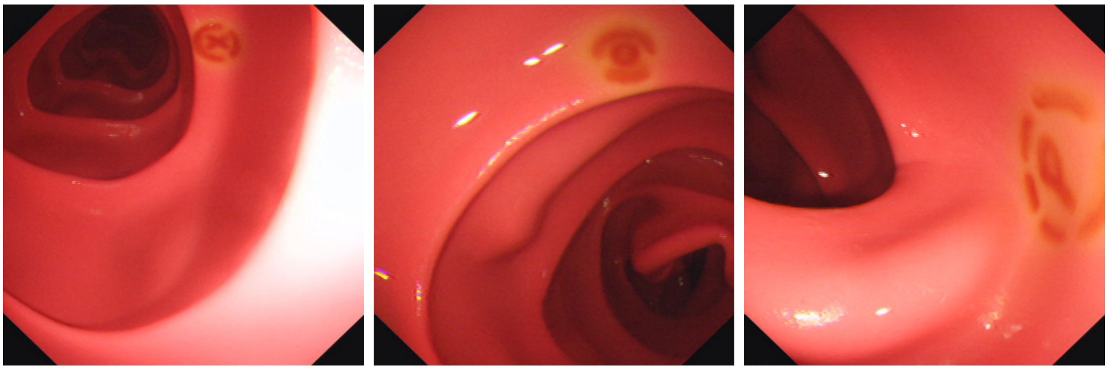


Figure 5.2: Examples of markers placed inside the colon model.

- Time measurements.

Screening Task. All participating endoscopists were given an instruction sheet on their task, which was to perform a screening colonoscopy on the model. Instead of polyps or other lesions, the model contained 15 markers - unique alphabetic characters in random order surrounded by circles. Figure 5.2 shows examples of such markers. The participants knew about the type of markers, while the number of markers was not disclosed. Furthermore, the participants knew they were going to be assessed anonymously. The instruction sheet can be found in Appendix C. The participants had an assistant available to turn the model to its sides or administer abdominal pressure, if necessary.

Data Collection. The actual data collection took place in two different endoscopy units. Before the actual procedure, we asked the participants for information on their status and experience and for their permission to associate their polyp detection rates and caecal intubation rates with the data. The polyp detection rate is measured as the fraction of patients in which polyps were found. Caecal intubation rate is the fraction of procedures in which the caecum was intubated (i.e. insertion was completed). Each participant was introduced to the system setup individually and was encouraged to read the instructions carefully. Any open questions were discussed to a point that the endoscopist appeared to have a clear understanding of the task. Before the actual screening procedure, each participant completed a short training phase to get accustomed to the colonoscopy model. The training was stopped once the endoscope was advanced past the rectum.

The participants then performed the screening procedure without supervision, saying out loud the particular letter whenever they found a marker. This gave us verbal confirmation that a visualised marker was actually recognised.

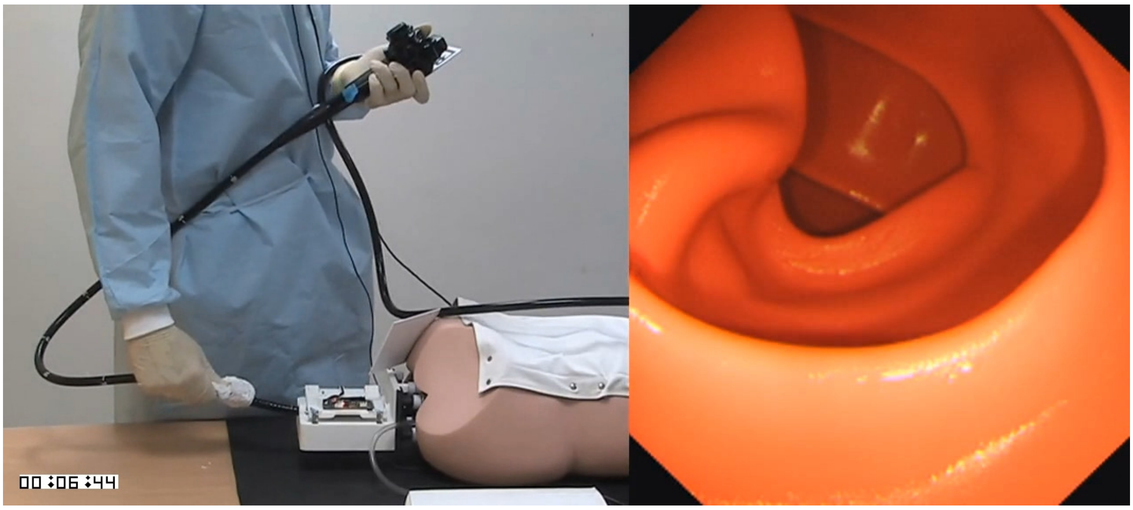


Figure 5.3: Frame from one of the videos presented to the assessors.

The feedback we received after the procedure was mixed. Many considered the system to be a realistic simulation of a colonoscopy procedure, while others criticised the lack of suction and the different feel of the resin material compared to real colonic mucosa.

A total of 32 endoscopy procedures were performed, out of which 4 had to be discarded due to difficulties with the technical setup. Those procedures were not repeated because this may have skewed the data due to learning effects.

5.2.2 Assessment Phase

For the offline assessment of the procedures the recordings had to be presented in a way that closely resembles being in the room during the procedure. We therefore placed the video of the endoscope camera side-by-side with the external view of the whole scene with the endoscopist and training model in the field of view. Figure 5.3 shows an example frame from one of the procedures. Prior to assessment, both experts performed a procedure on the training model, in order to be able to better interpret the videos with respect to the different handling experienced with the model. They were then shown 3 examples of videos for which they were asked to reach a consensus on their rating. After this they received the videos of the actual procedures together with the assessment forms to complete on their own.

As described earlier, the full DOPS assessment form contains 4 groups of criteria: 1) Assessment, Consent, Communication, 2) Safety and Sedation, 3) Endoscopic Skills During Insertion and Procedure, and 4) Diagnostic and Therapeutic Ability. The actual technical skills for performing a high quality colonoscopy are mostly in the third and fourth group. Since our objective is to assess quality and skill from the endoscopic video and handling data alone, we left out all criteria which do not directly involve endoscope handling or interpretation of the endoscopic visualisation. We arranged the remaining criteria into a condensed assessment form, with the original grade descriptors placed beside the rating scales. The criteria in the form are (for detailed descriptions of the criteria see Section 3.2):

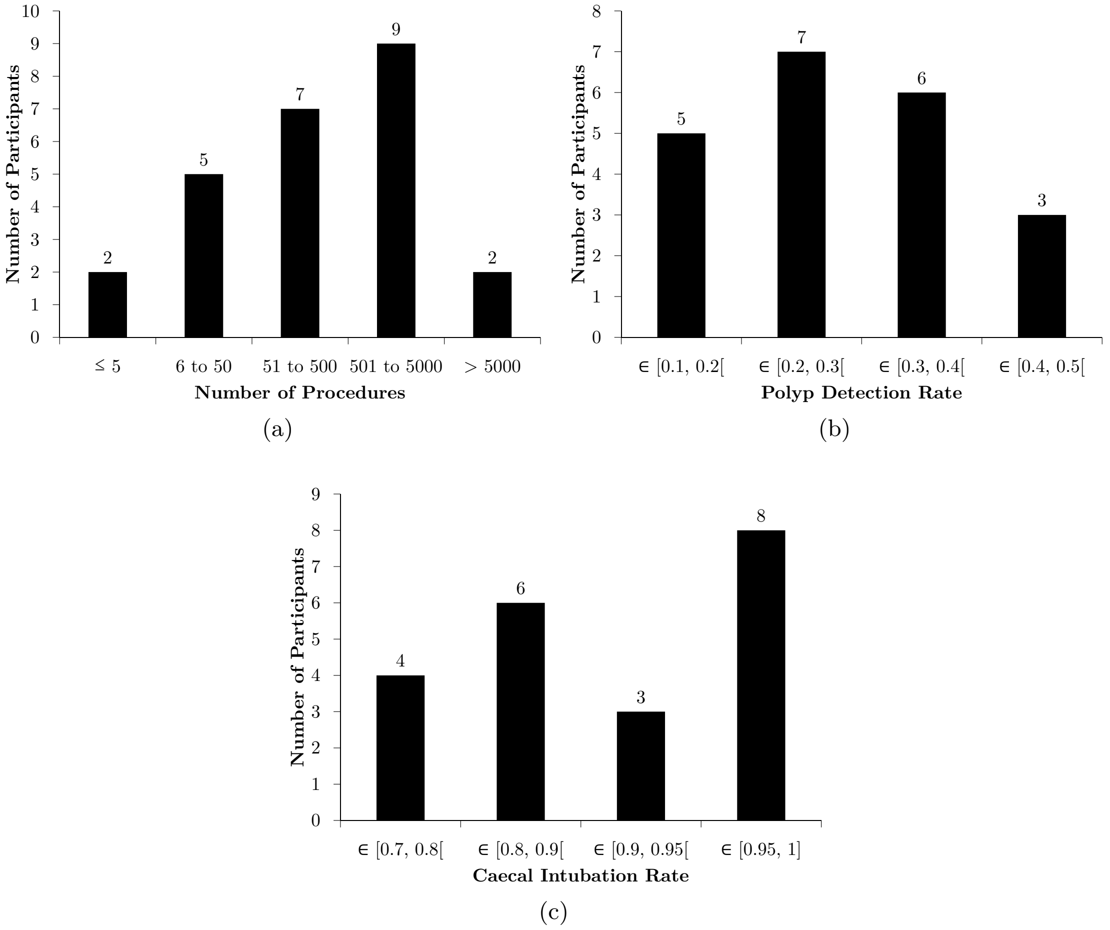
- Maintains luminal view / inserts in luminal direction (abbreviated *Lumen* in the following discussion)
- Uses torque steering and control knobs appropriately (*Handling*)
- Recognises and logically resolves loop formation (*Looping*)
- Uses position change and abdominal pressure to aid luminal views (*Position and pressure*)
- Completes procedure in reasonable time (*Time*)
- Adequate mucosal visualisation (*Visualisation*)
- Performance in insertion phase (*Insertion performance*)
- Performance in withdrawal phase (*Withdrawal performance*)
- Whole procedure performance (*Procedure performance*)

The last 3 criteria are the additional summary scores we have described in Section 3.2. All criteria had to be rated on an integer scale from 1 to 4 as in the original DOPS assessment form. *Completes procedure in reasonable time* was the only exception. As nothing in the grade descriptors associated with the JAG DOPS form justified a rating of 4, we only allowed a scale from 1 to 3. The complete assessment form can be found in Appendix D.

5.3 Data Analysis

The data collection resulted in a data set from 28 colonoscopy procedures with corresponding expert assessments. Among the endoscopists were a total of 6 consultants, 18 registrars, 2 SHOs (senior house officers, i.e., junior doctors) and 2 nurses, from 2 different hospitals.

Out of the 28 recorded procedures, 3 were incomplete due to the endoscopists not being able to advance further. These procedures differ significantly from the completed ones, as they all show prolonged attempts to advance the endoscope in the early stages of insertion,



(a)

(b)

(c)

Figure 5.4: Frequency distributions for the recorded experience and performance data of the participating endoscopists (a) number of self-reported colonoscopy procedures, (b) polyp detection rate, (c) caecal intubation rate.

followed by rapid withdrawal, largely omitting screening. We chose to exclude these procedures from further analysis due to these very different characteristics. Detection and assessment of incomplete procedures is, without doubt, at least as important as the assessment of complete procedures. The limited availability of data, however, forced us to leave this problem to future research. We will therefore, from here on, analyse the 25 complete procedures. In the following, we list the collected data and provide visual representations of their frequency distributions.

Endoscopist Data. The experience of the endoscopists according to the number of procedures performed is shown in the frequency distribution in Figure 5.4(a).

Unfortunately, the contributing endoscopy suites do not routinely record the adenoma detection rates of their endoscopists. However, the polyp detection rates (PDR) and caecal intubation rates (CIR) are available for the endoscopists who already have performed self-reported procedures. Frequency distributions of the PDR and CIR are shown in Figure 5.4(b,c).

Procedure Data. From the video data, we manually extracted the following basic procedure information:

- Insertion time (Time duration from anal intubation to commencement of withdrawal)
- Withdrawal time (Time duration from commencement of withdrawal to full withdrawal of the endoscope)
- Number of alphabetic markers detected (confirmed by the endoscopist saying the found letter on the audio recording)

Figure 5.5 shows frequency distributions of these measures.

Of the overall 15 alphabetic markers the participants found an average of 10.2. None of the participants detected all 15 markers. The fact that each single marker was detected by at least 10 of the participants, suggests that none of the markers was too difficultly placed to be detected. While having a small sample of only 25 procedures, this result supports the recent findings of significant miss rates in lesion detection.

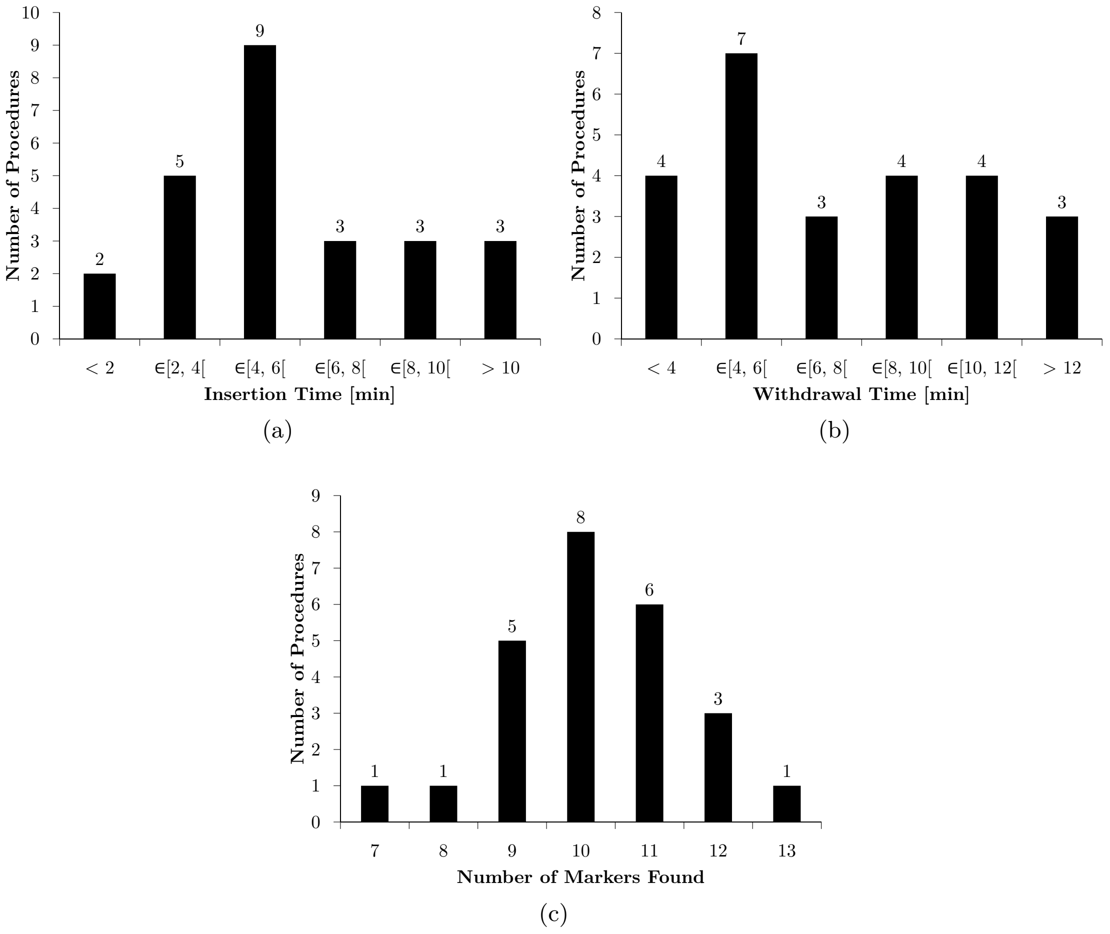
The average withdrawal time of 8.5 minutes can be considered adequate. A very recent guideline of the American College of Gastroenterology (ACG) recommends a mean withdrawal time of at least 6 minutes [3].

Expert Assessment. Both experts assessed all 25 complete procedures from the mentioned video recordings according to the assessment form. The resulting frequency distributions can be seen in Figures 5.6 and 5.7 for the DOPS criteria and the summary criteria, respectively.

The charts show that the two assessors have rated the procedures quite differently. In the following section we describe this tendency in more detail as we analyse inter-rater reliability.

5.3.1 Inter-Rater Reliability

Inter-rater reliability (or inter-rater/inter-coder agreement) describes the degree to which a number of independently working raters agree on the values of variables they are assessing. There is a wide variety of measures used for the assessment of inter-rater reliability



(a)

(b)

(c)

Figure 5.5: Frequency distributions for the procedure characteristics obtained from the endoscopic video data: (a) insertion time, (b) withdrawal time, (c) number of detected markers.

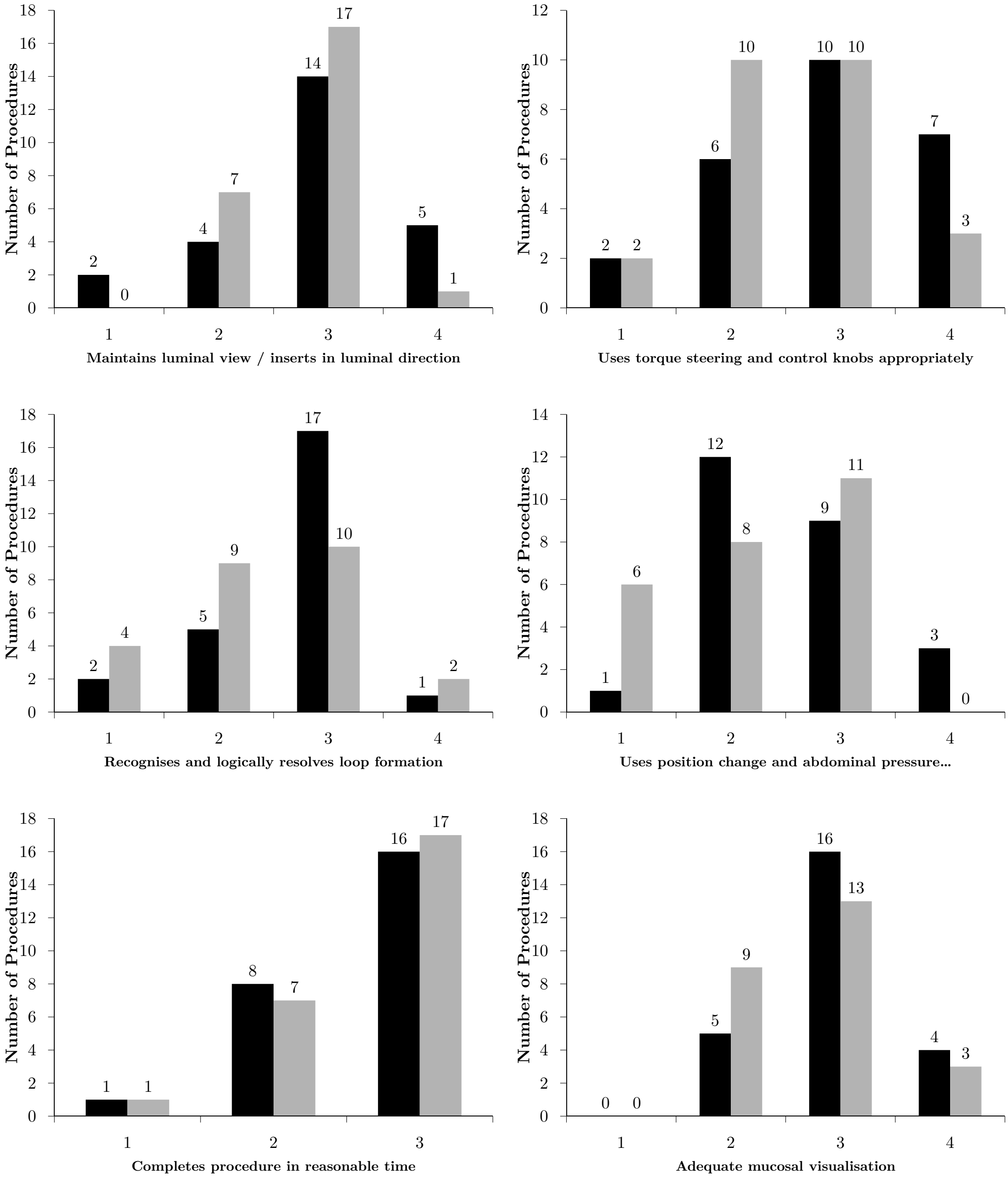


Figure 5.6: Frequency distributions for the DOPS criteria in the assessment form for assessor 1 (black bars) and assessor 2 (grey bars).

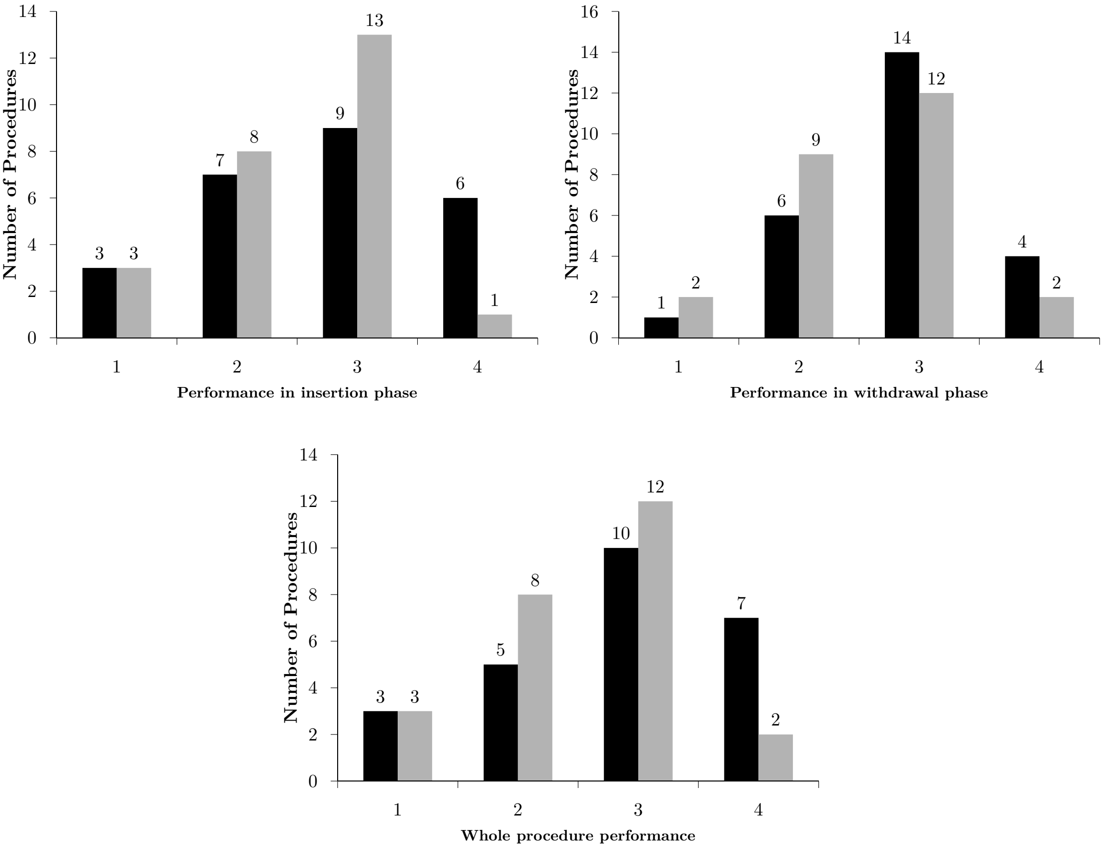


Figure 5.7: Frequency distribution for the summary criteria in the assessment form for assessor 1 (black bars) and assessor 2 (grey bars).

Table 5.1: Krippendorff's α coefficient for the rating criteria with 95% confidence intervals.

Criterion	α	Lower Limit 95% CI	Upper Limit 95% CI
Luminal View	0.23	-0.36	0.68
Handling	0.31	-0.07	0.64
Loops	0.47	0.18	0.72
Position and Pressure	0.32	-0.22	0.70
Time	0.05	-0.39	0.45
Visualisation	0.21	-0.36	0.64
Insertion Performance	0.37	0.07	0.61
Withdrawal Performance	0.34	-0.13	0.73
Procedure Performance	0.41	0.07	0.71

(see, e.g., [155, 156, 157]). We choose the widely used and recommended Krippendorff α coefficient for our analysis [158, 159]. α is a universally applicable, chance-corrected measure of inter-rater reliability, meaning that it takes into account the agreement that may occur by chance and corrects for it. It is, among others, applicable to ordinal data, which makes it an appropriate choice for our purposes. In its most general form, Krippendorff's α is defined as

$$\alpha = 1 - \frac{D_o}{D_e}, \tag{5.1}$$

with D_o and D_e being the observed and expected disagreement, respectively. It can take on values between -1 and 1. $\alpha = 1$ means that there is perfect agreement, $\alpha = 0$ that agreement is as good as chance agreement, and $\alpha < 0$ that there is systematic disagreement. Krippendorff [158] suggests to rely only on variables with a value of $\alpha > 0.8$, and recommends a minimum of $\alpha = 0.667$ to draw tentative conclusions from data. For a detailed description on Krippendorff's α and its underlying assumptions, please see Appendix A.

For easier interpretability, we complement α with the percentage of agreement between raters, and the percentages of approximate agreement split by the amount of error (e.g., raters agreed fully in 50% of the cases, disagreement was ≤ 1 on the rating scale in 75% of the cases, ≤ 2 in 95% of the cases, etc.).

Table 5.1 shows Krippendorff's α with 95% confidence intervals for the 9 assessed quality criteria. The achieved values for α are disappointing throughout, with only 3 criteria having an α coefficient significantly different from 0 (*Loops*, *Insertion Performance* and *Procedure Performance*), and none of the criteria fulfilling the minimum of $\alpha = 0.667$ to draw tentative conclusions. For comparison, Table 5.2 shows to what degree

Table 5.2: Degrees of agreement between the raters

Criterion	Full Agreement	Disagreement ≤ 1	Disagreement ≤ 2
Luminal View	76%	92%	100%
Handling	60%	88%	100%
Loops	64%	96%	100%
Position and Pressure	56%	96%	96%
Time	80%	96%	100%
Visualisation	72%	92%	100%
Insertion Performance	60%	92%	100%
Withdrawal Performance	72%	88%	100%
Procedure Performance	64%	88%	100%

the assessments differed between the assessors. On average, the assessors agreed fully on 67.1% of their ratings. A disagreement of greater than 1 level occurred only for the *position and pressure* criterion, and only for one of the 25 procedures.

Discussion

With the small sample size, we can only obtain a very rough estimate of inter-rater reliability, and the confidence intervals of α reflect this. Nevertheless, we expected better results, as great care has been taken in designing the experiment and instructing the assessors. While the feedback from the participants did not suggest any problems, it is possible that the necessary deviations from the original DOPS protocol (performing the procedures on a training model and assessing performance from video data) may have introduced a certain variability to the data. The agreement values in Table 5.2, however, encourage the interpretation that there may be sufficient agreement between the raters for our purposes. We may not be able to fully answer the question, to what degree DOPS assessment can be automated, but will attempt to at least find tendencies that point in this direction.

Future studies will seek to remove the restrictions of using a colonoscopy training model and offline video assessment, by designing a motion sensor that can be used in real colonoscopy procedures. The sample size and number of assessors in this study was restricted by limited resources. Future studies will require this limitation to be removed.

5.3.2 Associations Between Variables

Among all the measured variables, we consider the number of detected markers to be the one with highest validity in terms of measuring the quality of a procedure in our experimental setup. In a real colonoscopy procedure, finding lesions is the actual purpose.

In case a present lesion is missed, the procedure should have been done better, even if all measures indicate a high quality of screening. The situation is similar with the task of finding markers in our case. A good quality measure for colonoscopy procedures should therefore exhibit a strong association with the number of detected markers. The following analysis therefore begins with a view on the strength of association between endoscopist data, DOPS ratings and the number of detected markers, before highlighting a number of other interesting properties of the data set.

We report the strength of association between data using Kendall's τ statistic, and, for comparison, Pearson's product-moment correlation coefficient r. τ measures association of ranked data, and can be viewed as a non-parametric measure of statistical dependence. r measures the linear dependence between pairs of variables. For the exact definitions and underlying assumptions of these measures, please refer to Appendix A.

We consider τ to be the most suitable measure of association here, since it inherently regards the data as ordinal. While measures such as the withdrawal time or the number of detected markers may in themselves be interval data, we consider them here as measures of procedure quality with, if any, likely non-linear association. They should therefore be viewed as ordinal variables. While Pearson's r is generally only suitable for interval or ratio data, it is reported for comparison as it is commonly used in similar studies. We report the two statistics together with the associated p-values (two-tailed), as provided by the statistics software IBM SPSS 20.

Experience, Established Quality Measures and Number of Detected Markers. Looking at the available variables, one might expect strong associations between experience or skill level related variables, such as the number of procedures performed or the polyp detection rate, and the number of markers detected by the participants. Interestingly, the data suggests that these associations are, in fact, particularly weak. Table 5.3 shows the said measures when analysing the number of markers paired with the number of procedures, the polyp detection rate, the caecal intubation rate, and the manually assessed withdrawal time. In contrast with the other measures in Table 5.3, the withdrawal time does not reflect the experience or skill level of the endoscopist. We have added it because caecal intubation rate, adenoma detection rate (with is closely related to polyp detection rate) and withdrawal time have all been established as measures of procedure quality (for an overview see [160]).

With τ and r close to 0 and p-values clearly higher than the 0.1 significance level, none of the experience or skill level related variables can be associated with the number of detected markers. However, moderate association can be asserted in relation to the withdrawal time ($\tau = 0.367$, with $p = 0.017$). Since the layout of the training model was

Table 5.3: Measures of association between experience of the endoscopists, established quality measures and the number of detected markers.

	τ	r
No. of Procedures	-0.019 (p=0.90)	-0.083 (p=0.69)
Polyp Det. Rate	-0.209 (p=0.222)	-0.204 (p=0.38)
Caecal Intub. Rate	0.072 (p=0.68)	0.052 (p=0.82)
Withdrawal Time	0.367 (p=0.017)	0.47 (p=0.018)

the same for all participants, the reasons for the very low degree of association for the experience and skill level related variables must be sought elsewhere. One reason may be the deviations from real colonoscopy procedures, due to the use of a training model and the lack of suction functionality. However, we see no reason why this should have a stronger effect on experienced endoscopists than it has on beginners. A more likely scenario, in our opinion, is that the beginners' behaviour may be influenced heavily by the fact that they know they are being assessed (known as the Hawthorne effect [161, 162]), while experienced endoscopists may be less susceptible in this respect. The way the different hospitals measure polyp detection rate and caecal intubation rate may also have introduced variability. In the two endoscopy suites they are computed from the records of all colonoscopic procedures, whether mainly done for screening purposes or for therapeutic interventions. The fact that experienced endoscopists perform the more difficult therapeutic interventions, implies that the patient populations vary for endoscopists with different levels of experience. Furthermore, it is not recorded, whether the procedures were performed autonomously or with expert assistance. The moderate association between the withdrawal time and the number of detected markers supports the decisions of a number of professional bodies to suggest a minimal withdrawal time in their guidelines for colonoscopy procedures. Withdrawing the endoscope slowly allows a more careful examination of the mucosa and can therefore improve detection accuracy.

Expert Assessment and Number of Detected Markers. The criterion on the DOPS assessment form that is closest related to lesion detection is the quality of mucosal visualisation, for which one may expect the strongest association with the number of detected markers. Other possibly related measures are the appropriateness of procedure time and our additional summary measures for the performance in withdrawal phase and the whole procedure. Measures of association between these measures and the number of detected markers are shown in Tables 5.4 and 5.5 for assessors 1 and 2, respectively.

Table 5.4: Measures of association between expert ratings of assessor 1 and the number of detected markers.

Assessor 1	τ	r
Visualisation	0.065 (p=0.71)	0.060 (p=0.81)
Withdrawal Perf.	0.154 (p=0.37)	0.157 (p=0.49)
Overall Perf.	0.086 (p=0.61)	0.087 (p=0.71)
Proc. Time	0.052 (p=0.77)	0.053 (p=0.90)

Table 5.5: Measures of association between expert ratings of assessor 2 and the number of detected markers.

Assessor 2	τ	r
Visualisation	0.241 (p=0.17)	0.288 (p=0.18)
Withdrawal Perf.	0.284 (p=0.102)	0.329 (p=0.12)
Overall Perf.	0.279 (p=0.105)	0.314 (p=0.15)
Proc. Time	0.391 (p=0.030)	0.422 (p=0.045)

As suggested by the analysis of inter-rater reliability, we notice a significant difference between the two assessors. None of the ratings of assessor 1 can be associated with the number of detected markers, while for assessor 2 we can at least assert a moderate degree of association for one of the four reported ratings (precedure time with $\tau = 0.391$, $p = 0.03$) and a tendency towards association for two of the measures (with p-values just outside a significance level of p=0.1). Interestingly, no significant association can be asserted between the mucosal visualisation ratings of either assessor and the number of detected markers.

Experience and Expert Assessment. Another interesting aspect for analysis is whether the expert assessment is associated with endoscopist experience. Tables 5.6 and 5.7 provide an overview of this, listing all measures with significant association with at least one of the assessors. For all but one of the listed variables, the ratings of assessor 1 show moderate degrees of association with the number of self-reported procedures performed by the endoscopists, all statistically significant at the p=0.05 level. Associations are overall weaker using the ratings of assessor 2, with the exception of the criterion *ability to recognise and resolve loops*, which is the only instance where we can assert statistical significance at the p=0.05 level.

Table 5.6: Measures of association between expert ratings of assessor 1 and the number of self-reported procedures performed by the endoscopist.

Assessor 1	τ	r
Visualisation	0.364 (p=0.029)	0.253 (p=0.24)
Withdrawal Perf.	0.405 (p=0.014)	0.275 (p=0.17)
Overall Perf.	0.425 (p=0.008)	0.430 (p=0.024)
Proc. Time	0.345 (p=0.043)	0.261 (p=0.17)
Steering	0.412 (p=0.011)	0.408 (p=0.038)
Loops	0.248 (p=0.14)	0.177 (p=0.35)
Insertion Perf.	0.362 (p=0.024)	0.365 (p=0.069)

Table 5.7: Measures of association between expert ratings of assessor 2 and the number of self-reported procedures performed by the endoscopist.

Assessor 2	τ	r
Visualisation	0.266 (p=0.11)	0.088 (p=0.70)
Withdrawal Perf.	0.219 (p=0.18)	0.153 (p=0.47)
Overall Perf.	0.208 (p=0.20)	-0.035 (p=0.87)
Proc. Time	0.158 (p=0.35)	-0.026 (p=0.92)
Steering	0.304 (p=0.062)	0.154 (p=0.46)
Loops	0.551 (p=0.001)	0.373 (p=0.065)
Insertion Perf.	0.286 (p=0.082)	0.198 (p=0.36)

5.3.3 Discussion

Looking at these results, we can see that assessor 1 seems to rate in concordance with the experience of the endoscopists, while the ratings of assessor 2 agree more with the outcome of the procedure, i.e., the number of detected markers. This allows the interpretation that the two assessors have a different understanding of the criteria on the assessment form. It looks as if assessor 1 tended to value the skilful handling of an experienced endoscopist higher, while assessor 2 rewarded more the thoroughness of the screening, irrespective of handling skills. This could as well be one reason for the poor inter-rater reliability we found earlier.

For our objective of automatic assessment of DOPS criteria, we could either make a decision which interpretation of quality to value higher and choose the corresponding assessor, or we could train two different assessment models using the ratings of the two assessors separately. Since this is a decision where largely medical aspects have to be taken into account, we choose the second option and create separately trained assessment models. The following chapter will describe this approach in detail.

Chapter 6

Towards Automatic DOPS Assessment of Colonoscopy Procedures

6.1 Introduction

Given the data set obtained from the described experiment, we can compute the image measures proposed in Chapter 4 for each frame in the procedure videos. This results in an extensive resource of low level characteristics which we will use in this chapter to derive a large set of features describing various aspects of complete colonoscopy procedures. These include, among others, the insertion and withdrawal times, characteristics of endoscope handling and summary statistics of all image level measures.

Following the development of these procedure characteristics, we select the relevant ones for each of our chosen target DOPS and summary measures, before applying correlation analysis to further optimise the selection of features. Support vector regression models are then trained for mapping the procedure level measures to each of the chosen DOPS and summary measures. Given the low level of agreement between the two experts, we train and evaluate these models for each of the experts separately.

6.2 Automatic Measures of Characteristics of Colonoscopy Procedures

In Section 3.1 we have described a sensor prototype for measuring longitudinal and circular motion at the orifice. Having this data available for the 25 procedures, we now propose a number of measures for various procedure characteristics, which make use of the addition motion information. A selection of the proposed measures will be used as features in a machine learning approach to assessing DOPS criteria automatically. For most of these measures, the methods of measurement we propose can not be directly evaluated given our available data. At this point, the measures are propositions which need to be further analysed to prove their clinical value.

Many of the measures proposed in this section make use of the longitudinal and circular displacement, Δd_l and Δd_c, obtained from measurements of the orifice sensor. The sensor readings can be mapped onto a metric scale with knowledge of the sensor specifications, obtaining the metric displacements Δd_{lm} and Δd_{cm}. The orifice sensor performs measurement at a rate of $125\,Hz$ with a resolution of $800\,dpi$ (dots per inch). From the video perspective we are restricted to a frame rate of $20\,fps$. A given video contains a total of N frames. For frame $n \in \{1,..,N\}$ at time t_0 with respect to the previous frame $n-1$ at time $t_0 - 0.05\,s$, we compute the metric displacements as

$$\Delta d_{lm}(n) = \sum_{t=t_0-0.05s}^{t_0} \frac{\Delta d_l(t)}{800} \cdot 254\,\text{mm, and}$$

$$\Delta d_{cm}(n) = \sum_{t=t_0-0.05s}^{t_0} \frac{\Delta d_c(t)}{800} \cdot 254\text{mm.}$$

Following this conversion from the raw sensor readings to metric displacements, all further analysis will be performed at the time resolution constrained by the video frame rate. Numbers of frames can always be transformed to time durations by dividing the number of frames by the frame rate.

6.2.1 Estimating Depth of Insertion

Depth of insertion, in contrast to the metric displacement Δd_{lm}, measures, for every frame n, the length of the inserted part of the endoscope. Having this measure available would allow us to structure the procedures in a clinically meaningful way. We then could separate insertion from withdrawal, which is useful, as the requirements for endoscopic technique and handling are quite different in the two phases. An estimate of the depth of insertion also helps in assessing whether certain measures were maintained over the whole extent of the colon or whether in certain segments performance was weaker. Without these estimates, procedures can only be structured into time segments with no relation to the physical structure of the colon or the semantic structure of the procedure.

Assuming that no errors occur in measuring metric displacement Δd_{lm}, the insertion depth d_{lm} results from computing the cumulative sum over the metric displacement

$$d_{lm}(n) = \sum_{k=1}^{n} \Delta d_{lm}(k). \tag{6.1}$$

Performing this operation on our available data, the measurement error of the orifice sensor becomes apparent. The cumulative summing operation also leads to an accumulation of the error, meaning that d_{lm} does not return exactly to 0 after the endoscope has been fully withdrawn. Hence, before further analysing the procedure in terms of the insertion depth, we adjust d_m. Assuming the error is spread evenly over all N frames in the procedure, the adjusted insertion depth d^*_{lm} is computed as

$$d^*_{lm}(n) = \frac{n\, d_{lm}(N)}{N} d_{lm}(n)\,. \tag{6.2}$$

Methods for measuring characteristics derived from the insertion depth, such as the insertion and withdrawal times in the following section, make use of this adjusted measure.

6.2.2 Insertion and Withdrawal Time

We consider the insertion time T_I to be the time from intubation of the anus until commencement of withdrawal. The withdrawal time T_W then is the time from commencement of withdrawal until complete withdrawal of the endoscope. A simple approach to measuring the insertion and withdrawal times would be to find a time instance t_0 at which the endoscope was farthest inserted, such that $d^*_{lm}(t) \leq d^*_{lm}(t_0), \forall t \in [0, T_P]$, with T_P being the time taken for the whole procedure from intubation of the anus until complete withdrawal of the endoscope. The insertion and withdrawal times would then be computed as $T_I = t_0$ and $T_W = T_P - t_0$. To illustrate the problem with this approach, Figure 6.1 shows an example of the insertion depth computed for a complete procedure. Repeated insertion and withdrawal as in the figure between approximately $100\,s$ and $350\,s$ can be observed in many procedures. This can be due to an attempt to resolve a loop or get around a difficult bend. If it happens close to the caecum, the endoscope may be inserted farther than the actual extent of the colon. The insertion time measured using the point of farthest insertion would in this case be $245\,s$, which is significantly shorter than the correct insertion time of $485\,s$ (correct insertion and withdrawal times were assessed manually according to the above definitions).

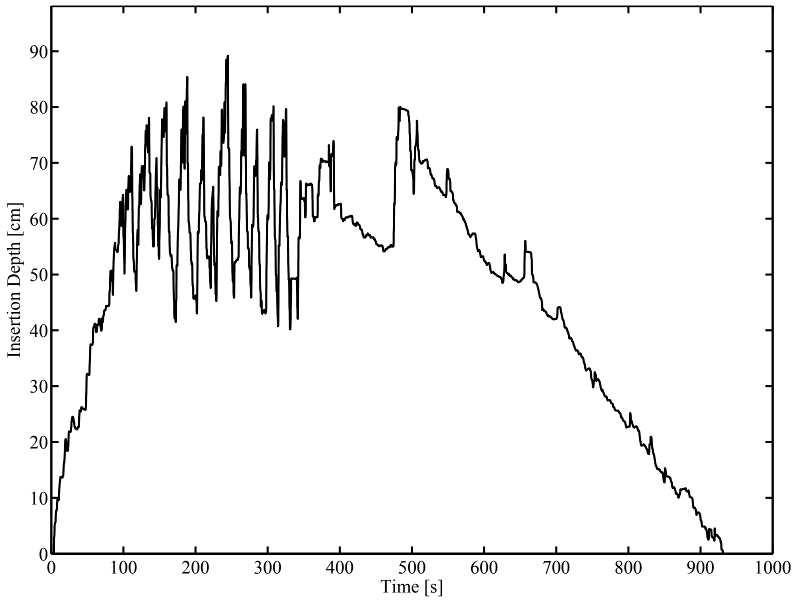


Figure 6.1: Insertion depth sensor reading without filtering.

In order to resolve this issue, we perform a moving average filtering on the insertion depth readings. For a window size of $2w + 1$ the resulting filtered insertion depth d_{lf} is computed as

$$d_{lf}(n) = \frac{1}{2w+1} \sum_{k=n-w}^{n+w} d^*_{lm}(k). \tag{6.3}$$

For a procedure video with N frames, in order to compute values for $n \leq w$ and $n > N-w$, we extend $d^*_{lm}(n)$ to $n < 1$ and $n > N$ in a symmetric fashion, such that $d^*_{lm}(1-k) = -d^*_{lm}(1+k)$ and $d^*_{lm}(N+k) = -d^*_{lm}(N-k)$ for $k \in \{1..w\}$. This padding ensures that insertion depth close to the orifice is not overestimated. Figure 6.2 shows the result of filtering the raw insertion depth in Figure 6.1 using a window of size $2w+1 = 271$ frames. Measuring the insertion time according the the method described above, replacing d^*_{lm} by d_{lf}, yields $T_I = 491\,s$ with an error of only $6\,s$ compared to the ground truth.

Evaluation. The proposed method was evaluated on the 25 procedure videos using leave-one-out cross-validation. Optimal window sizes $2w+1$ were learned from the training sets using an exhaustive search approach. The method was then applied to the test videos. The insertion time was estimated with a root mean squared (RMS) error of 6.47% of the correct insertion time, which corresponds to an absolute RMS error of $18.73\,s$. With respect to the correct withdrawal time, the estimated withdrawal time had an RMS error of 3.82%. The optimised window size amounted on average to $\overline{2w+1} = 273.8$ frames with a standard deviation of 16.2 frames.

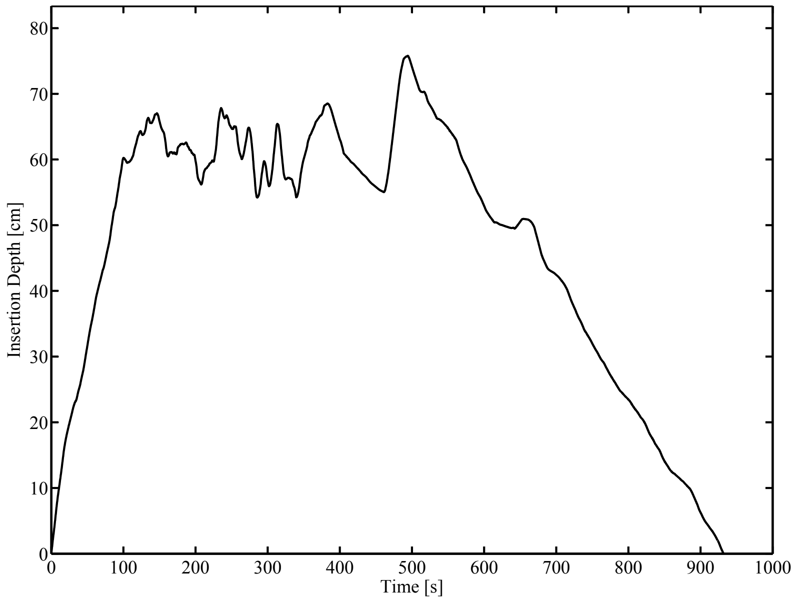


Figure 6.2: Insertion depth sensor reading using an optimal moving average filter.

Discussion. The filtering of the estimated insertion depth has an interpretation beyond the fact that it is useful for measuring insertion and withdrawal time. Repeated rapid pushing and pulling of the endoscope often coincides with repeated stretching and shortening of the colon without actually advancing past the current segment. However, the previously discussed segmentation of the colon according to the unfiltered insertion depth would indicate rapid movement between a number of segments. The filtering prevents this oversegmentation to a large degree. The filtered insertion depth is therefore preferable to the raw insertion depth for segmentation of the colon. The raw insertion depth, again, is more suitable when the handling of the endoscope is the object of analysis.

6.2.3 Summarising Measures over the Course of Procedures

The filtered insertion depth, defined above, allows us to divide the colon into segments of certain length. This segmentation, together with the metric longitudinal and circular displacement, enables us to define a number of novel measures of procedure characteristics, which shall later help us in the automatic measurement of DOPS criteria. In the following sections we use a number of methods to summarise measurements over the course of a procedure. Here, we introduce operators for each of these methods, which will simplify defining the large amount of proposed measures.

A complete procedure consists of N video frames, of which N_I frames lie within the insertion phase and $N_W = N - N_I$ frames remain for the withdrawal phase. This results in two sets $\mathcal{N}_I$ and $\mathcal{N}_W$, containing all frames from the insertion and withdrawal phases, respectively. The filtered insertion depth d_{lf}, discussed above, can be divided into K segments for both insertion and withdrawal phases. We define a segment simply as an interval of certain length on the domain of the insertion depth. For our analysis, we use a segment length of 5 cm, which we consider to be approximately the length of a colon segment that can be sufficiently examined without moving the camera. We can then define the sets $\mathcal{N}_{Ik}$ and $\mathcal{N}_{Wk}$ containing all frames n for which the insertion depth is in segment k during the insertion and withdrawal phases, respectively. The corresponding number of frames $N_{SI}(k)$, $N_{SW}(k)$ are the cardinalities of the sets $\mathcal{N}_{Ik}$ and $\mathcal{N}_{Wk}$. We can furthermore define the sets $\mathcal{K}_I$ and $\mathcal{K}_W$, containing all segments in the insertion and withdrawal phases, respectively. Throughout the next sections, we make use of the following operations:

For measures $M(n)$ defined for single frames n:

- arithmetic mean over the insertion phase:

$$\theta_I(M) = \frac{1}{N_I} \sum_{n \in \mathcal{N}_I} M(n),$$

- arithmetic mean over the withdrawal phase:

$$\theta_W(M) = \frac{1}{N_W} \sum_{n \in \mathcal{N}_W} M(n),$$

- standard deviation of $M(n)$ in the insertion/withdrawal phase:

$$s_P(M) = \sqrt{\frac{1}{N_P - 1} \sum_{n \in \mathcal{N}_P} (M(n) - \theta_P(M))^2}, \; P \in \{I, W\},$$

- arithmetic mean of $M(n)$ in the k-th insertion/withdrawal segment:

$$\rho_P(M, k) = \frac{1}{N_{SP}(k)} \sum_{n \in \mathcal{N}_{Pk}} M(n), \; P \in \{I, W\},$$

- standard deviation of $M(n)$ in the k-th insertion/withdrawal segment:

$$s_{SP}(M, k) = \sqrt{\frac{1}{N_{SP}(k) - 1} \sum_{n \in \mathcal{N}_{Pk}} (M(n) - \rho_P(M, k))^2}, \; P \in \{I, W\},$$

- arithmetic mean over the B highest valued frames $\hat{n}_b$, $\forall b \in \{1..B\}$, within the k-th segment:
$$\rho_{PB}(M,k) = \frac{1}{B}\sum_{\hat{n}_b} M(\hat{n}_b),\ P \in \{I,W\}.$$

For measures $M^*(k)$ defined as having one value per segment k:

- arithmetic mean over all segments in the insertion/withdrawal phase:
$$\overline{\theta}_P(M^*) = \frac{1}{K}\sum_{k\in\mathcal{K}_P} M^*(k),\ P \in \{I,W\},$$
- median over all segments in the insertion/withdrawal phase:
$$\widetilde{\theta}_P(M^*) = \widetilde{M^*}(k),\ \forall k \in \mathcal{K}_P,\ P \in \{I,W\},$$
with $\widetilde{\cdot}$ denoting the median operation.
- arithmetic mean over the N_H segments k_h with a value greater than the median value of the segments in the insertion/withdrawal phase:
$$\overline{\theta}_{PH}(M^*) = \frac{1}{N_H}\sum_{k_h} M^*(k_h),\ k_h \in \mathcal{K}_P,\ P \in \{I,W\}.$$
- arithmetic mean over the N_L segments k_l with a value smaller than the median value of the segments in the insertion/withdrawal phase:
$$\overline{\theta}_{PL}(M^*) = \frac{1}{N_L}\sum_{k_l} M^*(k_l),\ k_l \in \mathcal{K}_P,\ P \in \{I,W\}.$$

Which summarisation operation to choose depends on the properties of the measure in question. Details and reasons for the choice of a particular operation will be given together with the measure definitions in the following sections.

6.2.4 Measures of Time and Velocity

A number of important characteristics of a colonoscopy procedure can be determined from endoscope velocity and time durations. Insertion time and withdrawal time are examples of such characteristics. The withdrawal time has already been associated with procedure quality in a number of publications (see Section 2.2). Related characteristics may similarly be useful as quality indicators. In the following, we propose a number of measures based on the velocity of the endoscope and time durations.

Central Tendency and Variability of Time per Segment

We compute the time $T_{SI}(k)$ ($T_{SW}(k)$) spent in a segment k during insertion (withdrawal) from the number of frames during which the filtered insertion depth d_{lf} (6.3) is within the range of the segment in question. Using some of the previously introduced operators, we propose the following measures as alternatives to the plain insertion and withdrawal time:

- Average time per segment (insertion): $\overline{\theta}_I(T_{SI})$
- Average time per segment (withdrawal): $\overline{\theta}_W(T_{SW})$
- Median time per segment (insertion): $\widetilde{\theta}_I(T_{SI})$
- Median time per segment (withdrawal): $\widetilde{\theta}_W(T_{SW})$

These measures are closely related to the insertion and withdrawal time, while allowing for better comparison given colons of different length. In our experiment, the length of the colon is fixed, which makes insertion/withdrawal time and the average measures equivalent.

With regard to the quality of the procedure, the time taken for the different segments may yet hold more relevant information than can be expressed with a simple average or median. For example, if an endoscopist has spent on average $10\,s$ per segment on withdrawal, he may still have passed a number of segments very rapidly, possibly missing significant lesions. Therefore, we suggest additional measures reflecting the variability of the segment times. In our later analysis we use the following measures, while many other variants of these may be sensible:

- Average of segment times shorter than the median (insertion): $\overline{\theta}_{IL}(T_{SI})$,
- Average of segment times shorter than the median (withdrawal): $\overline{\theta}_{WL}(T_{SW})$.

Measures of Endoscope Handling Speeds

The longitudinal speed v_l of the shaft of the endoscope with respect to the sensor can be computed from the longitudinal metric displacement, by dividing it for any given frame by the duration of one frame, i.e.,

$$v_l(n) = \frac{\Delta d_{lm}}{0.05\,s}. \tag{6.4}$$

The speed v_c along the other axis, resulting from a rotation of the shaft of the endoscope, is computed accordingly from the circular metric displacement Δd_{cm}.

To the best of our knowledge, it has not yet been analysed whether there exists an association between the endoscope handling speeds and quality aspects of the procedure. Nevertheless, handling speed and the use of torque are mentioned as relevant characteristics in the texts on colonoscopy screening technique we have reviewed in Section3.3.

As with the time related measures, many different measures may be sensible for summarising the handling speeds over the course of the procedure. These include averages over segments or durations and different measures of variability. We treat insertion (forward) and withdrawal (backward) speeds (v_{lf}, v_{lb}) separately, as well as clockwise ($-$) and counter-clockwise ($+$) speeds (v_c^-, v_c^+), defining the sequences

$$v_{lf}(n) = \begin{cases} v_l(n) & n \in \{n : v_l(n) > 0\} \\ 0 & otherwise, \end{cases} \tag{6.5}$$

$$v_{lb}(n) = \begin{cases} -v_l(n) & n \in \{n : v_l(n) < 0\} \\ 0 & otherwise, \end{cases} \tag{6.6}$$

$$v_c^+(n) = \begin{cases} v_c(n) & n \in \{n : v_c(n) > 0\} \\ 0 & otherwise, \end{cases} \tag{6.7}$$

$$v_c^-(n) = \begin{cases} -v_c(n) & n \in \{n : v_c(n) < 0\} \\ 0 & otherwise. \end{cases} \tag{6.8}$$

We propose the following measures to summarise speeds over the course of a procedure:

- Average forward speed (insertion): $\theta_I(v_{lf})$
- Average backward speed (insertion): $\theta_I(v_{lb})$
- Average absolute speed through rotation (insertion): $\theta_I(|v_c|)$
- Average forward speed (withdrawal): $\theta_W(v_{lf})$
- Average backward speed (withdrawal): $\theta_W(v_{lb})$
- Average absolute speed through rotation (withdrawal): $\theta_W(|v_c|)$

Similar to to time averages per segment computed above, we also propose to measure these 6 characteristics by averaging the per-segment averages. This results in 6 more measures: $\overline{\theta}_I(\rho_I(v_{lf}))$, $\overline{\theta}_I(\rho_I(v_{lb}))$, $\overline{\theta}_I(\rho_I(|v_c|))$, $\overline{\theta}_W(\rho_W(v_{lf}))$, $\overline{\theta}_W(\rho_W(v_{lb}))$ and $\overline{\theta}_W(\rho_W(|v_c|))$. This gives us two different measures of central tendency for our later analysis.

By taking the average of the faster half of the segments during insertion and withdrawal, as it was done for the time per segment above, we get different representations of the handling velocities: $\overline{\theta}_{IH}(\rho_I(v_{lf}))$, $\overline{\theta}_{IH}(\rho_{IH}(v_{lb}))$, $\overline{\theta}_{IH}(\rho_I(|v_c|))$, $\overline{\theta}_{WH}(\rho_W(v_{lf}))$,

$\overline{\theta}_{WH}(\rho_W(v_{lb}))$ and $\overline{\theta}_{WH}(\rho_W(|v_c|))$. These measures capture more the tendency of the endoscopist to perform fast actions, which, on insertion, may be dangerous to the patient, and, on withdrawal, may lead to missed lesions.

6.2.5 Summarising Image Based Measures

In Chapter 4, we proposed a number of techniques for the measurement of characteristics of single images from colonoscopy procedures. For the analysis of the whole procedure, we need to summarise these measures and their behaviour over the course of the procedure.

The image based measures we have available are the clarity of the field of view $c(n)$, the presence of the lumen $l_{pr}(n)$ (and the unthresholded SVM output $l_{prp}(n)$), the quality of the luminal view $l_q(n)$, the distance to the next bend $db(n)$ and the 2-element lumen position vector $\mathbf{l_{pos}}(n)$. We also use a single dimensional representation of the lumen position, in the form of the distance to the image centre $l_{posD}(n)$, and the separate components $l_{posX}(n)$ and $l_{posY}(n)$.

In contrast to the other measures, the lumen presence is a binary measure. Hence, we need a different method for summarisation than we used for the other measures. We compute the fraction of frames in which the lumen is present for the insertion phase, the withdrawal phase and each segment during insertion and withdrawal. Furthermore, we propose a measure where this fraction is computed for the best B frames within a segment. Using this it is possible to ask questions such as, e.g., whether the lumen was present for at least 2 seconds during each segment. This results in 6 additional measures:

- Lumen presence fraction (insertion phase):
  $l_{prI} = N_{prI}/N_I$, with N_{prI} being the number of frames in which the lumen is present during the insertion phase.

- Lumen presence fraction (withdrawal phase):
  $l_{prW} = N_{prW}/N_W$, with N_{prW} being the number of frames in which the lumen is present during the withdrawal phase.

- Lumen presence fraction within segment k during insertion:
  $l_{prSI}(k) = N_{prSI}(k)/N_{SI}(k)$, with N_{prSI} being the number of frames in which the lumen is present during segment k in the insertion phase.

- Lumen presence fraction within segment k during insertion:
  $l_{prSW}(k) = N_{prSW}(k)/N_{SW}(k)$, with N_{prSW} being the number of frames in which the lumen is present during segment k in the withdrawal phase.

- Lumen presence fraction within the B best frames of segment k during insertion: $l_{prSIB}(k) = N_{prSIB}(k)/B$, with N_{prSIB} being the number of frames in which the lumen is present in the B best frames of segment k during insertion.
- Lumen presence fraction within the B best frames of segment k during withdrawal: $l_{prSWB}(k) = N_{prSWB}(k)/B$, with N_{prSWB} being the number of frames in which the lumen is present in the B best frames of segment k during withdrawal.

All other image based measures are summarised using most of the operations defined earlier in Section 6.2.3. Due to the resulting high number of measures, we leave the actual definitions to the end of this section, where a complete list of all proposed measures is shown.

6.2.6 Pushing Without a Clear View

During a colonoscopy procedure, one should only advance the endoscope when the field of view is clear enough to be able to tell the direction to advance to. We have discussed this in Section 3.3. Failure to comply with this directive can be detected when the sensor readings suggest a pushing attempt and our measure c for the clarity of the field of view (see Section 4.3) is low. Instead of measuring this on the basis of detecting events of pushing without clear view, we propose to compute the average of the forward speed v_{lf} (6.5) of the endoscope shaft during frames with low clarity. As per our definition of the clarity scale in Section 4.3, in images with a clarity of 1, the field of view is blurred to a degree that the position of the endoscope can at most be vaguely guessed. Our proposed measure of pushing without clear view P is therefore computed as:

$$P = \frac{1}{N_{lc}} \sum_{n \in \mathcal{N}_{lc}} v_{lf}(n), \tag{6.9}$$

with $\mathcal{N}_{lc} = \{n : c(n) < 1.5\}$, N_{lc} being the number of elements in $\mathcal{N}_{lc}$, and $c(n)$ being the earlier proposed automatic measure of clarity of the field of view (see Section 4.3). It can be divided into P_I and P_W for separate assessment of insertion and withdrawal phase by restricting n to the frames in the insertion and withdrawal phase, respectively.

6.2.7 Stationary Periods During Endoscope Insertion

During the insertion phase, the endoscopist may have difficulties advancing the endoscope due to situations such as challenging bends in the colon or the development of a loop, among others. Detection of such stationary periods may be helpful in the analysis of colonoscopy procedures. The presence of long stationary periods during insertion may indicate a difficult procedure or shortcomings in the technique of the endoscopist. Analysis

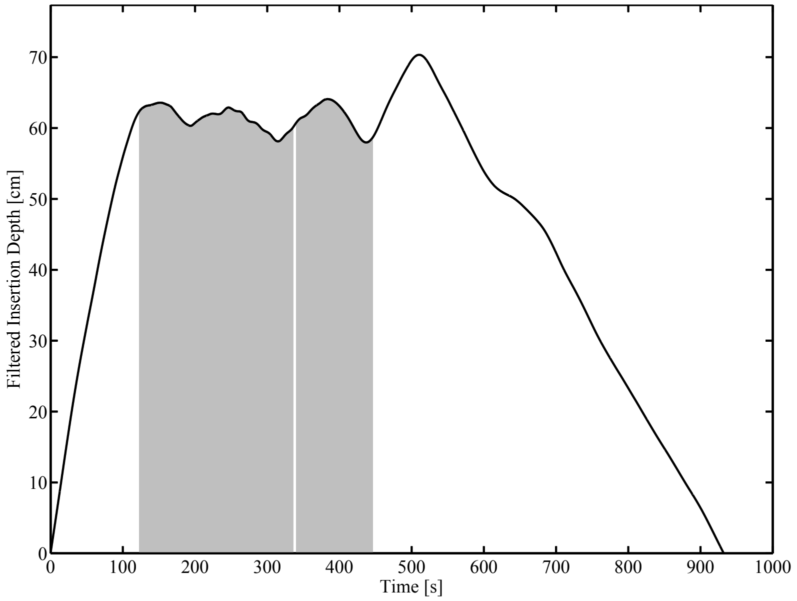


Figure 6.3: Filtered depth of insertion after filtering with a moving average filter of length $2w + 1 = 701$ frames. The grey areas mark the periods detected as stationary by our proposed method.

of the handling patterns during such periods may help assess the ability of the endoscopist to resolve problems such as loops. On withdrawal, stationary periods can occur due to close examination of found lesions.

We propose a measure for stationary periods, which makes use of a moving average filter as we used it already in Section 6.2.2 for measuring insertion and withdrawal time. Here, we use a second, even longer, moving average filter ($2w + 1 = 701$ frames) on the filtered insertion depth d_{lf} (6.3), in order to fully suppress most of the short-term forward and backward motion. Figure 6.3 shows the output d_{lff} of this moving average filter on the example that was shown earlier in Section 6.2.2 in connection with the measurement of insertion and withdrawal time. The actual stationary periods are found by thresholding the backward difference Δd_{lff}. We consider the procedure to be stationary if $-4 < \Delta d_{lff} < 4$. If Δd_{lff} is outside the $[-4, 4]$ interval, inbetween stationary regions, for a duration shorter than $1\,s$ at a time, it is considered stationary as well. This is to reduce the effect of outliers. In case Δd_{lff} is stationary over a duration of at least $10\,s$, we consider it a stationary period. The grey areas in Figure 6.3 mark the detected stationary periods. We propose the total time spent in stationary periods during insertion, T_s, as another quality related measure, which may help in assessing the endoscopists ability to handle difficult situations.

6.2.8 Loop Resolution

The ability to recognise and resolve loops is a major criterion in the DOPS assessment form. It is therefore important to include a measure of loop resolution into our analysis. A characteristic handling pattern associated with attempts of loop resolution is to twist the endoscope clockwise while pulling back. We use the stationary periods defined above to only count loop resolution attempts that occur during stationarity. We consider clockwise withdrawal to take place, if, during such a stationary period, longitudinal backward speed $v_{lb} > 1\,cm/s$ and clockwise rotation speed $v_c^- > 0.5\,cm/s$. Sporadic clockwise withdrawal, resulting from measurement errors or an unintended endoscope motion, is removed by a morphological opening operation using a structuring element of length 5. We sum up the duration of clockwise withdrawal motions during the insertion phase to obtain the total clockwise withdrawal duration T_l.

For our later analysis, we set this duration into relation with the total stationary time, using the measure $L = T_l/T_s$ as a feature. As stated earlier, stationary periods can occur not only in connection with the presence of a loop. The proposed measure is therefore clearly not optimal for measuring loop resolution attempts. Future studies will need to apply camera motion estimation to improve the detection of loops. The total duration of stationary periods could then be replaced by the duration where loops where actually present.

6.2.9 Discussion and Overview of All Proposed Measures

Most of the measures we have proposed in this section are unvalidated propositions which will need to be further analysed to prove their clinical value. The main purpose for the measures in this thesis is to characterise colonoscopy procedures as comprehensively as possible. In the next section we will analyse whether this characterisation allows us to train an automatic rating system able to assess colonoscopy procedures according to the chosen target criteria from the JAG DOPS assessment system. To summarise this section, we list all the proposed measures for procedure characteristics in Tables 6.1 and 6.2. These sets of measures form the basis of an approach to measuring DOPS quality criteria we describe in the following section.

Table 6.1: List of all proposed measures for characteristics of the insertion phase of colonoscopy procedures.

1	T_I	Insertion time
2	$\overline{\theta}_I(T_{SI})$	Average time taken per segment for insertion phase
3	$\overline{\theta}_{IL}(T_{SI})$	Average of segment times shorter than the median for insertion phase
4	$\theta_I(l_q)$	Luminal view quality averaged over insertion frames
5	$\overline{\theta}_I(\rho_I(l_q))$	Average luminal view quality per insertion segment
6	$\overline{\theta}_{IL}(\rho_I(l_q))$	Average luminal view quality for segments below the median
7	$\overline{\theta}_I(\rho_{I50}(l_q))$	Average luminal view quality per segment counting only the best 50 frames per segment
8	$\overline{\theta}_{IL}(\rho_{I50}(l_q))$	Average luminal view quality for segments below the median counting only the best 50 frames per segment
9	$\theta_I(db)$	Distance to next bend averaged over insertion frames
10	$\overline{\theta}_I(\rho_I(db))$	Average distance to next bend per insertion segment
11	l_{prI}	Fraction of frames in which the lumen was present
12	$\overline{\theta}_I(l_{prSI})$	Average fraction of frames per segment in which the lumen was present
13	$\overline{\theta}_{IL}(l_{prSI})$	Average lumen presence fraction for segments below the median
14	$\overline{\theta}_I(l_{prSI50})$	Average lumen presence fraction per segment (counting only the best 50 frames per segment)
15	$\overline{\theta}_{IL}(l_{prSI50})$	Average lumen presence fraction for segments below the median (counting only the best 50 frames per segment)
16	$\theta_I(l_{prp})$	Unthresholded SVM output of the lumen presence averaged over insertion frames
17	$\overline{\theta}_I(\rho_I(l_{prp}))$	Average unthresholded SVM output of the lumen presence per insertion segment
18	$\overline{\theta}_{IL}(\rho_I(l_{prp}))$	Average unthresholded SVM output of the lumen presence for segments below the median
19	$\overline{\theta}_I(\rho_{I50}(l_{prp}))$	Average unthresholded SVM output of the lumen presence, counting only the best 50 frames per segment
20	$\overline{\theta}_{IL}(\rho_I 50(l_{prp}))$	Average unthresholded SVM output of the lumen presence for segments below the median, counting only the best 50 frames per segment
21	$\theta_I(c)$	Clarity averaged over insertion frames
22	$\overline{\theta}_I(\rho_I(c))$	Average clarity per insertion segment

23	$\overline{\theta}_{IL}(\rho_I(c))$	Average clarity for segments below the median
24	$\overline{\theta}_I(\rho_{I50}(c))$	Average clarity per segment, counting only the best 50 frames per segment
25	$\overline{\theta}_{IL}(\rho_I 50(c))$	Average clarity for segments below the median, counting only the best 50 frames per segment
26	P_I	Measure for blind pushing during insertion
27	$\theta_I(l_{posD})$	Distance from the lumen position to the centre of the image averaged over insertion frames
28	$\overline{\theta}_I(\rho_I(l_{posD}))$	Average distance from the lumen position to the centre of the image per insertion segment
29	$\theta_I(l_{posX})$	Lumen location in X direction averaged over insertion frames
30	$\overline{\theta}_I(\rho_I(l_{posX}))$	Average lumen location in X direction per insertion segment
31	$\theta_I(l_{posY})$	Lumen location in Y direction averaged over insertion frames
32	$\overline{\theta}_I(\rho_I(l_{posY}))$	Average lumen location in Y direction per insertion segment
33	$s_I(l_{posX})$	Standard deviation of lumen location in X direction during insertion
34	$\overline{\theta}_I(s_{SI}(l_{posX}))$	Average standard deviation of lumen location in X direction per insertion segment
35	$s_I(l_{posY})$	Standard deviation of lumen location in Y direction during insertion
36	$\overline{\theta}_I(s_{SI}(l_{posY}))$	Average standard deviation of lumen location in Y direction per insertion segment
37	$\theta_I(v_{lf})$	Longitudinal forward speed averaged over insertion frames
38	$\overline{\theta}_I(\rho_I(v_{lf}))$	Average longitudinal forward speed per insertion segment
39	$\overline{\theta}_{IL}(\rho_I(v_{lf}))$	Average longitudinal forward speed for segments below the median
40	$\theta_I(v_{lb})$	Longitudinal backward speed averaged over insertion frames
41	$\overline{\theta}_I(\rho_I(v_{lb}))$	Average longitudinal backward speed per insertion segment
42	$\overline{\theta}_{IL}(\rho_I(v_{lb}))$	Average longitudinal backward speed for segments below the median
43	$\theta_I(\|v_c\|)$	Absolute circular speed averaged over insertion frames
44	$\overline{\theta}_I(\rho_I(\|v_c\|))$	Average absolute circular speed per insertion segment
45	$\overline{\theta}_{IL}(\rho_I(\|v_c\|))$	Average absolute circular speed for segments below the median
46	T_s	Time spent in stationary periods during insertion
47	L	Measure of attempted loop resolution in stationary periods during insertion

Table 6.2: List of all proposed measures for characteristics of the withdrawal phase of colonoscopy procedures.

48	T_W	Withdrawal time
49	$\overline{\theta}_W(T_{SW})$	Average time taken per segment for withdrawal phase
50	$\overline{\theta}_{WL}(T_{SW})$	Average of segment times shorter than the median for withdrawal phase
51	$\theta_W(l_q)$	Luminal view quality averaged over withdrawal frames
52	$\overline{\theta}_W(\rho_W(l_q))$	Average luminal view quality per withdrawal segment
53	$\overline{\theta}_{WL}(\rho_W(l_q))$	Average luminal view quality for segments below the median
54	$\overline{\theta}_W(\rho_{W50}(l_q))$	Average luminal view quality per segment, counting only the best 50 frames per segment
55	$\overline{\theta}_{WL}(\rho_{W50}(l_q))$	Average luminal view quality for segments below the median, counting only the best 50 frames per segment
56	$\theta_W(db)$	Distance to next bend averaged over withdrawal frames
57	$\overline{\theta}_W(\rho_W(db))$	Average distance to next bend per withdrawal segment
58	l_{prW}	Fraction of frames in which the lumen was present
59	$\overline{\theta}_W(l_{prSW})$	Average fraction of frames per segment in which the lumen was present
60	$\overline{\theta}_{WL}(l_{prSW})$	Average lumen presence fraction for segments below the median
61	$\overline{\theta}_W(l_{prSW50})$	Average lumen presence fraction per segment (counting only the best 50 frames per segment)
62	$\overline{\theta}_{WL}(l_{prSW50})$	Average lumen presence fraction for segments below the median (counting only the best 50 frames per segment)
63	$\theta_W(l_{prp})$	Unthresholded SVM output of the lumen presence averaged over withdrawal frames
64	$\overline{\theta}_W(\rho_W(l_{prp}))$	Average unthresholded SVM output of the lumen presence per withdrawal segment
65	$\overline{\theta}_{WL}(\rho_W(l_{prp}))$	Average unthresholded SVM output of the lumen presence for segments below the median
66	$\overline{\theta}_W(\rho_{W50}(l_{prp}))$	Average unthresholded SVM output of the lumen presence, counting only the best 50 frames per segment
67	$\overline{\theta}_{WL}(\rho_{W50}(l_{prp}))$	Average unthresholded SVM output of the lumen presence for segments below the median, counting only the best 50 frames per segment

68	$\theta_W(c)$	Clarity averaged over withdrawal frames
69	$\overline{\theta}_W(\rho_W(c))$	Average clarity per withdrawal segment
70	$\overline{\theta}_{WL}(\rho_W(c))$	Average clarity for segments below the median
71	$\overline{\theta}_W(\rho_{W50}(c))$	Average clarity per segment, counting only the best 50 frames per segment
72	$\overline{\theta}_{WL}(\rho_{W50}(c))$	Average clarity for segments below the median, counting only the best 50 frames per segment
73	P_W	Measure for blind pushing during withdrawal
74	$\theta_W(l_{posD})$	Distance from the lumen position to the centre of the image averaged over withdrawal frames
75	$\overline{\theta}_W(\rho_W(l_{posD}))$	Average distance from the lumen position to the centre of the image per withdrawal segment
76	$\theta_W(l_{posX})$	Lumen location in X direction averaged over withdrawal frames
77	$\overline{\theta}_W(\rho_W(l_{posX}))$	Average lumen location in X direction per withdrawal segment
78	$\theta_W(l_{posY})$	Lumen location in Y direction averaged over withdrawal frames
79	$\overline{\theta}_W(\rho_W(l_{posY}))$	Average lumen location in Y direction per withdrawal segment
80	$s_W(l_{posX})$	Standard deviation of lumen location in X direction during withdrawal
81	$\overline{\theta}_W(s_{SW}(l_{posX}))$	Average standard deviation of lumen location in X direction per withdrawal segment
82	$s_W(l_{posY})$	Standard deviation of lumen location in Y direction during withdrawal
83	$\overline{\theta}_W(s_{SW}(l_{posY}))$	Average standard deviation of lumen location in Y direction per withdrawal segment
84	$\theta_W(v_{lf})$	Longitudinal forward speed averaged over withdrawal frames
85	$\overline{\theta}_W(\rho_W(v_{lf}))$	Average longitudinal forward speed per withdrawal segment
86	$\overline{\theta}_{WH}(\rho_W(v_{lf}))$	Average longitudinal forward speed for segments above the median
87	$\theta_W(v_{lb})$	Longitudinal backward speed averaged over withdrawal frames
88	$\overline{\theta}_W(\rho_W(v_{lb}))$	Average longitudinal backward speed per withdrawal segment
89	$\overline{\theta}_{WH}(\rho_W(v_{lb}))$	Average longitudinal backward speed for segments above the median
90	$\theta_W(\|v_c\|)$	Absolute circular speed averaged over withdrawal frames
91	$\overline{\theta}_W(\rho_W(\|v_c\|))$	Average absolute circular speed per withdrawal segment
92	$\overline{\theta}_{WL}(\rho_W(\|v_c\|))$	Average absolute circular speed for segments below the median

6.3 Inferring DOPS Quality Criteria from Measures of Procedure Characteristics

In Section 2.2.2, we have introduced the JAG DOPS assessment form as a way of measuring the quality of individual colonoscopy procedures. This form of assessment is preferable to the long term statistics over many procedures, which are routinely used in clinical practice. However, the fact that two assessors need to be present for the assessment of each procedure makes this approach impractical for routine use. In Section 3.2, we have identified a subset of the criteria in the JAG DOPS assessment form, which may be automatically measurable given the video data and sensor recordings of the procedure. In the following, we propose an approach to inferring these criteria from the measures of procedure characteristics introduced in the previous section.

6.3.1 Objective

Although each of the 92 measures has been chosen carefully to reflect a certain characteristic of a colonoscopy procedure, some of them may be irrelevant with respect to some DOPS criteria. We also have used different summarisation operations for the same underlying measure, meaning that there may be many redundant features. The objective is, therefore, to find, for each DOPS criterion, a combination of measures that describe the criterion accurately, and given this combination, to train a predictive model that is able to accurately assess unseen procedures.

6.3.2 Method

A major difficulty is the size of the data set we have available. 25 examples are insufficient for performing automatic feature selection, as we did within the machine learning framework presented in Chapter 4. Having 92 features available this method would lead to overfitting and, therefore, poor performance on unseen data. Choosing the relevant features manually based on domain knowledge is not trivial, since there is seldom an intuitive advantage of any summarisation operation over the others for the different features.

We therefore take a hybrid approach to feature selection. We organise the features into groups, each of which is made up of a set of features representing the same underlying procedure characteristic. We then use our domain knowledge to select the relevant groups (characteristics) for each of the target DOPS measures. Within each group we perform correlation analysis to chose the most relevant feature of each group. Only the feature

Table 6.3: Groups of features and their relevance for the different DOPS criteria.

	Lumen	**Handling**	**Looping**	**Time**	**Visualisation**	**Insertion Perf.**	**Withdrawal Perf.**	**Overall Perf.**
INSERTION								
Backward Speed	-	×	-	×	-	×	-	×
Forward speed	-	×	-	×	-	×	-	×
Circular Speed	-	×	-	-	-	×	-	×
Blind Pushing	-	×	×	-	-	×	-	×
Clarity	×	×	-	-	-	×	-	×
Loop Resolution	-	×	×	-	-	×	-	×
Lumen Dist. To Centre	×	×	-	-	-	×	-	×
Dist. To Bend	×	-	-	-	-	×	-	×
Lumen Pos. X	×	×	-	-	-	×	-	×
Lumen Pos. X Stdev	×	×	-	-	-	×	-	×
Lumen Pos. Y	×	×	-	-	-	×	-	×
Lumen Pos. Y Stdev	×	×	-	-	-	×	-	×
Lumen Presence	×	×	-	-	-	×	-	×
Lumen View Quality	×	×	-	-	-	×	-	×
Stationary Time	-	×	×	-	-	×	-	×
Time	-	×	×	×	-	×	-	×
WITHDRAWAL								
Backward Speed	-	-	-	×	×	-	×	×
Forward speed	-	-	-	×	×	-	×	×
Circular Speed	-	×	-	-	×	-	×	×
Blind Pushing	-	×	×	-	-	-	×	×
Clarity	×	×	-	-	×	-	×	×
Lumen Dist. To Centre	×	×	-	-	×	-	×	×
Dist. To Bend	×	-	-	-	×	-	×	×
Lumen Pos. X	×	×	-	-	×	-	×	×
Lumen Pos. X Stdev	×	×	-	-	×	-	×	×
Lumen Pos. Y	×	×	-	-	×	-	×	×
Lumen Pos. Y Stdev	×	×	-	-	×	-	×	×
Lumen Presence	×	×	-	-	×	-	×	×
Lumen View Quality	×	×	-	-	×	-	×	×
Time	-	-	-	×	×	-	×	×

with the highest correlation with the measure in the training examples is retained in each group. This approach allows us to reduce the dimensionality of the feature space significantly, which, in turn, should increase the performance of the learning algorithm and lead to better predictive models. Table 6.3 lists the target measures and feature groups we identified and shows which group we consider relevant for each of the target features. By using this method, depending on the DOPS criterion in question, the number of features is reduced to between 5 and 30.

Similarly as for the ordinal measures in Chapter 4, we use ν support vector regression models for prediction of the DOPS criteria. To make best use of the small data set, the SVM is trained in a nested cross-validation scheme. In the outer loop, in each cross-validation fold, we leave out a single example for testing and hand the rest to the inner loop. In the inner loop, parameters are optimised with a grid search approach, again using leave-one-out cross-validation.

6.3.3 Evaluation

In Section 5.3, we have discussed the results of our data collection involving 2 trained assessors. We have shown that the assessors differed significantly in their ratings, possibly due to conflicting interpretations of the JAG DOPS criteria. It appears that assessor 1 tended to value the skilful handling of an experienced endoscopist higher, while assessor 2 rewarded more the thoroughness of the screening, irrespective of handling skills. We consider it therefore reasonable to use the ratings of the two assessors separately and evaluate our system once for assessor 1 and once for assessor 2.

Table 6.4: Agreement between predictions of the automatic system and the ratings of assessor 1, measured using Kendall's τ, Pearson's r and Krippendorff's α. For τ and r the values in parentheses are the corresponding p-values. For α the table shows the lower and upper limits of the 95% confidence intervals in parentheses.

	τ (p-value)	α (95% CI lower lim., upper lim.)	r (p-value)
Lumen	0.442 (0.006)	0.169 (-0.111, 0.426)	0.648 (<0.001)
Handling	0.516 (0.001)	0.642 (0.412, 0.815)	0.705 (<0.001)
Looping	0.294 (0.077)	0.276 (-0.056, 0.564)	0.465 (0.019)
Time	0.290 (0.088)	0.229 (-0.154, 0.538)	0.344 (0.092)
Visualisation	0.460 (0.005)	0.607 (0.499, 0.711)	0.541 (0.005)
Insertion	0.404 (0.011)	0.434 (0.239, 0.618)	0.525 (0.007)
Withdrawal	0.488 (0.003)	0.566 (0.325, 0.756)	0.490 (0.013)
Overall	0.586 (<0.001)	0.400 (0.208, 0.575)	0.783 (<0.001)

Table 6.5: Agreement between predictions of the automatic system and the ratings of assessor 2, measured using Kendall's τ, Pearson's r and Krippendorff's α. For τ and r the values in parentheses are the corresponding p-values. For α the table shows the lower and upper limits of the 95% confidence intervals in parentheses.

	τ (p-value)	α (95% CI lower lim., upper lim.)	r (p-value)
Lumen	-0.459 (0.007)	-0.530 (-1.000, -0.052)	-0.573 (0.003)
Handling	0.241 (0.138)	0.213 (-0.054, 0.445)	0.335 (0.102)
Looping	-0.016 (0.940)	0.024 (-0.316, 0.328)	0.093 (0.660)
Time	-0.005 (1.000)	-0.202 (-0.660, 0.209)	-0.491 (0.013)
Visualisation	0.081 (0.639)	0.131 (-0.333, 0.525)	0.120 (0.567)
Insertion	-0.263 (0.109)	-0.216 (-0.664, 0.137)	-0.431 (0.031)
Withdrawal	-0.041 (0.818)	0.026 (-0.309, 0.322)	-0.171 (0.414)
Overall	0.081 (0.629)	0.095 (-0.239, 0.377)	-0.105 (0.618)

Table 6.6: Agreement between the ratings of assessor 1 and the ratings of assessor 2, measured using Kendall's τ, Pearson's r and Krippendorff's α. For τ and r the values in parentheses are the corresponding p-values. For α the table shows the lower and upper limits of the 95% confidence intervals in parentheses.

	τ (p-value)	α (95% CI lower lim., upper lim.)	r (p-value)
Lumen	0.217 (0.262)	0.230 (-0.360, 0.680)	0.314 (0.126)
Handling	0.311 (0.084)	0.310 (-0.070, 0.640)	0.366 (0.072)
Looping	0.494 (0.008)	0.470 (0.180, 0.720)	0.571 (0.003)
Time	0.041 (0.864)	0.050 (-0.390, 0.450)	0.051 (0.810)
Visualisation	0.185 (0.342)	0.210 (-0.360, 0.640)	0.181 (0.387)
Insertion	0.347 (0.055)	0.370 (0.070, 0.610)	0.462 (0.020)
Withdrawal	0.323 (0.082)	0.340 (-0.130, 0.730)	0.381 (0.060)
Overall	0.402 (0.025)	0.410 (0.070, 0.710)	0.518 (0.008)

For the performance evaluation of our proposed method, we compute the strength of association between the trained SVMs and each of the two assessors. We are using Kendall's τ statistic for this analysis (see Appendix A), again providing Pearson's r for comparison. The results are shown in Tables 6.4 and 6.5. For comparison with the earlier inter-rater reliability analysis, we report also Krippendorff's α coefficient, and Table 6.6 shows the agreement between the two assessors.

We can see a moderate association for most of the measures when comparing to assessor 1. Interestingly, the method fails to achieve any significant agreement with assessor 2. This may be due to inconsistent rating or the use of criteria we have failed to encode in our features. In any way, this again supports the finding that the two assessors rated differently. We see that the association between our method and assessor 1 is superior to the association between the ratings of the two assessors in all except the *looping* criterion. Association between the predictions of our method and the ratings of assessor 1 is always statistically significant at the 0.1 significance level, whereas between the two assessors, the association is insignificant for 3 criteria.

6.3.4 Discussion

The results our method achieves when compared to assessor 1 are promising. A significant degree of association for all criteria indicates that certain criteria can indeed be measured automatically from video and motion sensor data. The poor performance when compared to assessor 2, however, suggests that the data we collected suffers from inconsistencies. We don't have enough data examples and assessors to investigate the issue sufficiently. The problem will have to be addressed with a larger scale experiment, which we were unable to conduct given our available resources.

Chapter 7

Conclusions

7.1 Summary

In this thesis, we have taken a novel approach to assessing quality in colonoscopy procedures. We investigated, to what degree an established system for the subjective assessment of individual procedures (JAG DOPS) can be automated. In the course of this research, we have developed an automated assessment system which uses combinations of image and video analysis methods and machine learning algorithms. As the building blocks for this system, we have proposed a large number of lower-level measures for quality related characteristics in images and complete procedures. The development of these measures was oriented towards the overall objective of assessing DOPS criteria, while care was taken to make each of the measures stand on its own as a representation of relevant image or procedure characteristics.

For single images, we proposed a novel measure for the clarity of the field of view, extending the current state of the art by introducing multiple grades of clarity as opposed to the previously proposed binary classification into informative and indistinct images. By introducing measures for different characteristics of luminal views in single images, i.e., luminal view quality, lumen presence, position of the lumen and distance to the closest bend, we achieve a detailed description of colonoscopic images with direct implications to visualisation quality and endoscope handling skills. The image measures are based on models which were trained using a universal machine learning framework involving automatic feature selection and different variants of support vector machines. We proposed a number of novel features for these models. For the image clarity measure, we computed different representations of the amount of structure in the image based on wavelet decomposition of the image and distances between intensity histograms of lines in the image. The lumen characteristics are based on intensity, shape and colour features obtained from maximally stable extremal regions (MSER) in the image. Furthermore, all image based measures benefit from the proposed novel methods for detecting and inpainting specular highlights in endoscopic images. The methods for computing these image based measures achieved promising results.

For the characterisation of complete procedures we proposed to incorporate measurements of a motion sensor, located outside the anus, which measures longitudinal and circular displacement of the endoscope. Due to the motion sensor being optimised for a specific colonoscopy training model, it was necessary to design an experiment for data collection, in which video and sensor data could be recorded simultaneously. The obtained data was complemented by information on the experience of the participating endoscopists and assessments of the procedures by two trained experts according to JAG DOPS criteria. This way we were able to collect video and motion sensor data, for the development of measures for procedure characteristics, and associated DOPS ratings to train and evaluate models for automatic DOPS assessment.

Given the obtained motion sensor data, we presented a method to estimate the depth of insertion of the endoscope from these readings. The speed of the endoscope can be directly obtained. We used these measurements in methods to automatically infer the insertion and withdrawal times of procedures and to detect stationary periods during insertion, attempts of loop resolution and occurrences of pushing without a clear view. The estimated depth of insertion also allowed us to divide the colon into a number of segments, opening up new possibilities for summarising the behaviour of certain characteristics over the course of a procedure. Without this the only way was to analyse the behaviour over time, lacking any form of spatial information. Combining the spatial segmentation with the time scale, handling patterns and the set of proposed image measures, we created a set of 92 procedure measures.

As a final contribution we introduced a method to infer a selection of JAG DOPS criteria and summary scores from these procedure measures. The method involved feature selection, based on domain knowledge and correlation analysis, and the training of support vector regression models. The interpretation of the achieved results is not straightforward, as the ratings of the two experts differed significantly. We took the approach to use the ratings of the two experts separately, create two different sets of assessment models, and then analyse the achieved results separately. Using the data of assessor 1, moderate to high agreement could be achieved, while the system was unable to learn useful assessment models from the data of assessor 2.

7.2 Discussion and Future Directions

The results achieved by the proposed automatic assessment system suggest that there is a potential for automatic systems to routinely assess colonoscopy procedures. Video data together with motion sensor measurements appear to be sufficient to measure a selection of criteria from the JAG DOPS assessment system automatically. However, due to the variation in the performance of the assessment models depending on which assessor is chosen as a reference, the results remain inconclusive. The reasons for the poor agreement between the two assessors may be the deviations from the protocol of the original DOPS assessment or the failure to achieve a consensus between the assessors and their interpretations of the DOPS criteria. This issue needs to be investigated in a larger study involving a higher number of procedures and assessors, ideally with an experimental setup that closer resembles the original DOPS protocol. In this research programme, our resources did not permit extending the study any further. Nevertheless, the fact that the agreement with assessor 1 was significant for the majority of target measures (and stronger than the agreement between the assessors in a number of cases) is a promising result.

The proposed measures for image and procedure characteristics were designed to have clinical relevance on their own, describing the procedures in commonly used terms and using combinations of low-level features with clear interpretations. Feature selection was intentionally used instead of transformation methods for dimensionality reduction in order to achieve this. The proposed measures are therefore directly usable as objective statistics in automatically generated procedure reports. Furthermore, after computational optimisations, they could be presented to the endoscopist in real-time during the procedure. Further studies should investigate the potential of the proposed measures in similar applications.

Bibliography

[1] U.S. Preventive Services Task Force. Screening for colorectal cancer: U.S. preventive services task force recommendation statement. *Annals of Internal Medicine*, 149(9):627–637, Nov 2008.

[2] World Gastroenterology Organisation. World gastroenterology organisation/international digestive cancer alliance practice guidelines: Colorectal cancer screening. `http://www.worldgastroenterology.org/colorectal-cancer-screening.html`: [Last accessed: 23 Sep 2012], 2006.

[3] D. K. Rex, D. A. Johnson, J. C. Anderson, P. S. Schoenfeld, C. A. Burke, and J. M. Inadomi. American college of gastroenterology guidelines for colorectal cancer screening 2008. *The American Journal of Gastroenterology*, 104:739–750, 2009.

[4] V. S. Benson, J. Patnick, A. K. Davies, M. R. Nadel, R. A. Smith, and W. S. Atkin. Colorectal cancer screening: A comparison of 35 initiatives in 17 countries. *International Journal of Cancer*, 122(6):1357–1367, 2008.

[5] B. Bressler, L. Paszat, Z. Chen, D. Rothwell, C. Vinden, and L. Rabeneck. Rates of new or missed colorectal cancers after colonoscopy and their risk factors: a population-based analysis. *Gastroenterology*, 132(1):96–102, 2007.

[6] S. Stryker, B. Wolff, C. Culp, S. Libbe, D. Ilstrup, and R. MacCarty. Natural history of untreated colonic polyps. *Gastroenterology*, 93(5):1009–1013, 1987.

[7] J. D. Waye, D. K. Rex, and C. B. Williams, editors. *Colonoscopy: Principles and Practice, Second Edition.* Wiley-Blackwell, September 2009.

[8] H. S. Cooper. Intestinal neoplasms. In S. E. Mills, D. Carter, J. K. Greenson, V. E. Reuter, and M. H. Stoler, editors, *Sternberg's Diagnostic Surgical Pathology*, chapter 34, pages 1543–1601. Lippincott Williams & Wilkins, 2004.

[9] Participants in the Paris Workshop. The paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to december 1, 2002. *Gastrointestinal Endoscopy*, 58(6, Supplement 1):S3 – S43, 2003.

[10] J. Church. Adenoma detection rate and the quality of colonoscopy: The sword has two edges. *Diseases of the Colon & Rectum*, 51(5):520–523, May 2008.

[11] J. R. Barton, S. Corbett, and C. P. van der Vleuten. The validity and reliability of a direct observation of procedural skills assessment tool: assessing colonoscopic skills of senior endoscopists. *Gastrointestinal Endoscopy*, 75(3):591 – 597, 2012.

[12] M. Vassiliou, P. Kaneva, B. Poulose, B. Dunkin, J. Marks, R. Sadik, G. Sroka, M. Anvari, K. Thaler, G. Adrales, J. Hazey, J. Lightdale, V. Velanovich, L. Swanstrom, J. Mellinger, and G. Fried. Global assessment of gastrointestinal endoscopic skills (GAGES): a valid measurement tool for technical skills in flexible endoscopy. *Surgical Endoscopy*, 24(8):1834–1841, August 2010.

[13] World Health Organization. Media Centre Fact Sheet No. 297. `http://www.who.int/mediacentre/factsheets/fs297/en/index.html`: [Last accessed: 13 Nov 2011], 2011.

[14] B. Levin, D. A. Lieberman, B. McFarland, R. A. Smith, D. Brooks, K. S. Andrews, C. Dash, F. M. Giardiello, S. Glick, T. R. Levin, P. Pickhardt, D. K. Rex, A. Thorson, and S. J. Winawer. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the american cancer society, the US multi-society task force on colorectal cancer, and the american college of radiology. *CA: A Cancer Journal for Clinicians*, 58(3):130–160, 2008.

[15] J. Hunter and J. Sackier. *Minimally invasive surgery.* McGraw-Hill, New York, 1993.

[16] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain. Wireless capsule endoscopy. *Nature*, 405:417, 2000.

[17] A. Huang and R. M. Summers. Computer processing methods for virtual endoscopy. In I. Bankman, editor, *Handbook of Medical Image Processing and Analysis*, chapter 48, pages 833–845. Academic Press, 2008.

[18] R. Eliakim. Wireless capsule video endoscopy: Three years of experience. *World Journal of Gastroenterology*, 10(9):1238–1239, 2004.

[19] K. Schulmann, S. Hollerbach, K. Kraus, J. Willert, T. Vogel, G. Möslein, C. Pox, M. Reiser, A. Reinacher-Schick, and W. Schmiegel. Feasibility and diagnostic utility of video capsule endoscopy for the detection of small bowel polyps in patients with hereditary polyposis syndromes. *The American Journal of Gastroenterology*, 100(1):27, 2005.

[20] Z. Fireman, A. Glukhovsky, H. Jacob, A. Lavy, S. Lewkowicz, and E. Scapa. Wireless capsule endoscopy. *Israel Medical Association Journal*, 4:717–719, 2002.

[21] R. Eliakim, K. Yassin, Y. Niv, Y. Metzger, J. Lachter, E. Ga, B. Sapoznikov, F. Konikoff, G. Leichtmann, Z. Fireman, et al. Prospective multicenter performance evaluation of the second-generation colon capsule compared with colonoscopy. *Endoskopie heute*, 23(02):144–149, 2010.

[22] J. Liang, W. Higgins, R. Summers, and H. Yoshida. Introduction to the special section on virtual endoscopy. *IEEE Transactions on Medical Imaging*, 23(11):1333–1334, 2004.

[23] H. M. Fenlon, D. P. Nunes, P. D. Clarke, and J. T. Ferrucci. Colorectal neoplasm detection using virtual colonoscopy: a feasibility study. *Gut*, 43(6):806–811, 1998.

[24] C. Kay, D. Kulling, R. Hawes, J. Young, and P. Cotton. Virtual endoscopy-Comparison with colonoscopy in the detection of space-occupying lesions of the colon. *Endoscopy*, 32(3):226–232, 2000.

[25] M. Macari, A. Milano, M. Lavelle, P. Berman, and A. Megibow. Comparison of time-efficiency CT colonography with two- and three-dimensional colonic evaluation for detecting colorectal polyps. *AJR American Journal of Roentgenology*, 174(6):1543–1549, 2000.

[26] T. White, G. Avery, N. Kennan, A. Syed, J. Hartley, and J. Monson. Virtual colonoscopy vs conventional colonoscopy in patients at high risk of colorectal cancer–a prospective trial of 150 patients. *Colorectal Disease*, 11(2):138–145, 2009.

[27] Centers for Medicare & Medicaid Services (CMS). MLN matters number MM6578: Screening computed tomography colonography (CTC) for colorectal cancer. `http://www.cms.hhs.gov/MLNMattersArticles/downloads/MM6578.pdf`: [Last accessed: 24 Sep 2012].

[28] D. Lieberman. Quality and colonoscopy: a new imperative. *Gastrointestinal Endoscopy*, 61(3):392 – 394, 2005.

[29] J. M. Inadomi. In search of quality colonoscopy. *Gastroenterology*, 135(6):1845 – 1847, 2008.

[30] D. Rex, J. Bond, S. Winawer, T. Levin, R. Burt, D. Johnson, L. Kirk, S. Litlin, D. Lieberman, J. Waye, et al. Quality in the technical performance of colonoscopy and the continuous quality improvement process for colonoscopy: recommendations of the U. S. Multi-Society Task Force on Colorectal Cancer. *The American journal of gastroenterology*, 97(6):1296–1308, 2002.

[31] N. Baxter, M. Goldwasser, L. Paszat, R. Saskin, D. Urbach, and L. Rabeneck. Association of colonoscopy and death from colorectal cancer. *Annals of Internal Medicine*, 150(1):1, 2009.

[32] A. Pabby, R. Schoen, J. Weissfeld, R. Burt, J. Kikendall, P. Lance, M. Shike, E. Lanza, and A. Schatzkin. Analysis of colorectal cancer occurrence during surveillance colonoscopy in the dietary polyp prevention trial. *Gastrointestinal endoscopy*, 61(3):385–391, 2005.

[33] D. Rex, C. Cutler, G. Lemmel, E. Rahmani, D. Clark, D. Helper, G. Lehman, and D. Mark. Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies. *Gastroenterology*, 112(1):24 – 28, 1997.

[34] D. Heresbach, T. Barrioz, M. Lapalus, D. Coumaros, P. Bauret, P. Potier, D. Sautereau, C. Boustière, J. Grimaud, C. Barthélémy, et al. Miss rate for colorectal neoplastic polyps: a prospective multicenter study of back-to-back video colonoscopies. *Endoscopy*, 40(4):284, 2008.

[35] S. Thomas-Gibson, P. Rogers, S. Cooper, R. Man, M. Rutter, N. Suzuki, D. Swain, A. Thuraisingam, and W. Atkin. Judgement of the quality of bowel preparation at screening flexible sigmoidoscopy is associated with variability in adenoma detection rates. *Endoscopy*, 38(5):456–460, 2006.

[36] F. Froehlich, V. Wietlisbach, J.-J. Gonvers, B. Burnand, and J.-P. Vader. Impact of colonic cleansing on quality and diagnostic yield of colonoscopy: the european panel of appropriateness of gastrointestinal endoscopy european multicenter study. *Gastrointestinal Endoscopy*, 61(3):378 – 384, 2005.

[37] R. H. Lee, R. S. Tang, V. R. Muthusamy, S. B. Ho, N. K. Shah, L. Wetzel, A. S. Bain, E. E. Mackintosh, A. M. Paek, A. M. Crissien, L. J. Saraf, D. M. Kalmaz, and T. J. Savides. Quality of colonoscopy withdrawal technique and variability in adenoma detection rates (with videos). *Gastrointestinal Endoscopy*, 74(1):128 – 134, 2011.

[38] J. C. van Rijn, J. B. Reitsma, J. Stoker, P. M. Bossuyt, S. J. van Deventer, and E. Dekker. Polyp miss rate determined by tandem colonoscopy: A systematic review. *The American Journal of Gastroenterology*, 101(2):343–350, 2006.

[39] S. B. Ahn, D. S. Han, H. S. Cho, T. J. Byun, T. Y. Kim, C. S. Eun, Y. C. Jeon, and J. H. Sohn. Does colonoscopic withdrawal time affect polyp detection rate? *Gastrointestinal Endoscopy*, 69(5):AB231 – AB232, 2009. DDW Abstract Issue 2009, Digestive Disease Week 2009.

[40] R. Barclay, J. Vicari, A. Doughty, J. Johanson, and R. Greenlaw. Colonoscopic withdrawal times and adenoma detection during screening colonoscopy. *New England Journal of Medicine*, 355(24):2533, 2006.

[41] R. L. Barclay, J. J. Vicari, J. F. Johanson, and R. L. Greenlaw. Effect of a prespecified minimum colonoscopic withdrawal time on adenoma detection rates during screening colonoscopy. *Gastrointestinal Endoscopy*, 63(5):AB81, 2006.

[42] R. Gupta, M. Steinbach, K. V. Ballman, V. Kumar, and P. C. de Groen. Colorectal cancer despite colonoscopy: Critical is the endoscopist, not the withdrawal time. *Gastroenterology*, 136(5):A–55, May 2009.

[43] C. Almansa, M. Shahid, M. Heckman, S. Preissler, and M. Wallace. Association between visual gaze patterns and adenoma detection rate during colonoscopy: A preliminary investigation. *The American Journal of Gastroenterology*, 106(6):1070–1074, 2011.

[44] D. K. Rex, J. L. Petrini, T. H. Baron, A. Chak, J. Cohen, S. E. Deal, B. Hoffman, B. C. Jacobson, K. Mergener, B. T. Petersen, M. A. Safdi, D. O. Faigel, and I. M. Pike. Quality indicators for colonoscopy. *Gastrointestinal Endoscopy*, 63(4, Supplement 1):S16 – S28, 2006.

[45] American Society for Gastrointestinal Endoscopy. Quality improvement of gastrointestinal endoscopy: guidelines for clinical application. from the ASGE american society for gastrointestinal endoscopy. *Gastrointestinal Endoscopy*, 49(6):842–844, 1999.

[46] J. Johanson, C. Schmitt, T. Deas Jr, G. Eisen, M. Freeman, J. Goldstein, D. Jensen, D. Lieberman, S. Lo, A. Sahai, et al. Quality and outcomes assessment in Gastrointestinal Endoscopy. *Gastrointestinal Endoscopy*, 52(6 Pt 1):827, 2000.

[47] J. Rey, R. Lambert, et al. ESGE recommendations for quality control in gastrointestinal endoscopy: guidelines for image documentation in upper and lower GI endoscopy. *Endoscopy*, 33(10):901–903, 2001.

[48] D. K. Rex. Maximizing detection of adenomas and cancers during colonoscopy. *American Journal of Gastroenterology*, 101(12):2866–2877, December 2006.

[49] S. C. Chen and D. K. Rex. Endoscopist can be more powerful than age and male gender in predicting adenoma detection at colonoscopy. *American Journal of Gastroenterology*, 102(4):856–861, 2007.

[50] D. K. Rex. Colonoscopic withdrawal technique is associated with adenoma miss rates,. *Gastrointestinal Endoscopy*, 51(1):33 – 36, 2000.

[51] R. Barton. Validity and reliability of an accreditation assessment for colonoscopy. *Gut*, 57(Suppl. 1):A2, 2008.

[52] R. Sedlack. The mayo colonoscopy skills assessment tool: validation of a unique instrument to assess colonoscopy skills in trainees. *Gastrointestinal Endoscopy*, 72(6):1125–1133, 2010.

[53] M. Arnold, A. Ghosh, G. Lacey, S. Patchett, and H. Mulcahy. Indistinct frame detection in colonoscopy videos. In *Proc. 13th Int. Machine Vision and Image Processing Conf. IMVIP '09*, pages 47–52, 2009.

[54] J. Oh, S. Hwang, J. Lee, W. Tavanapong, J. Wong, and P. de Groen. Informative frame classification for endoscopy video. *Medical Image Analysis*, 11(2):110–127, 2007.

[55] T. P. Grantcharov, L. Carstensen, and S. Schulze. Objective assessment of gastrointestinal endoscopy skills using a virtual reality simulator. *Journal of the Society of Laparoendoscopic Surgeons*, 9:130–133(4), 2005.

[56] S. Y. Yi, K. H. Ryu, H. S. Woo, W. Ahn, W. S. Kim, and D. Y. Lee. Quantitative analysis of colonoscopy skills using the KAIST-Ewha colonoscopy simulator II. In *Proc. Frontiers in the Convergence of Bioscience and Information Technologies*, pages 519–524, 2007.

[57] Y. H. An, S. Hwang, J. Oh, J. Lee, W. Tavanapong, P. C. de Groen, and J. Wong. Informative-frame filtering in endoscopy videos. In J. M. Fitzpatrick and J. M. Reinhardt, editors, *Medical Imaging 2005: Image Processing*, volume 5747-1, pages 291–302. SPIE, 2005.

[58] Y. Cao, W. Tavanapong, D. Li, J. Oh, P. de Groen, and J. Wong. A Visual Model Approach for Parsing Colonoscopy Videos. In *Proc. 3rd International Conference on Image and Video Retrieval*, page 160. Springer Verlag, 2004.

[59] B.-L. Yeo and B. Liu. Rapid scene analysis on compressed video. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(6):533–544, December 1995.

[60] J. Oh, S. Hwang, W. Tavanapong, P. C. de Groen, and J. Wong. Blurry-frame detection and shot segmentation in colonoscopy videos. In *Proc. Storage and Retrieval Methods and Applications for Multimedia*, volume 5307-1, pages 531–542. SPIE, 2004.

[61] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, Nov. 1986.

[62] Y. Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):800–810, August 2001.

[63] C. B. Williams. Insertion technique. In J. D. Waye, D. K. Rex, and C. B. Williams, editors, *Colonoscopy: Principles and Practice*, chapter 29, pages 535–559. Wiley-Blackwell, 2009.

[64] S. Hwang, J. Oh, J. Lee, Y. Cao, W. Tavanapong, D. Liu, J. Wong, and P. de Groen. Automatic measurement of quality metrics for colonoscopy videos. In *Proc. 13th annual ACM international conference on Multimedia*, pages 912–921. ACM New York, NY, USA, 2005.

[65] D. Liu, Y. Cao, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Mining colonoscopy videos to measure quality of colonoscopic procedures. In *Proc. 5th IASTED International Conference: Biomedical Engineering*, pages 409–414, Anaheim, CA, USA, 2007. ACTA Press.

[66] D. Liu, Y. Cao, W. Tavanapong, J. Wong, J. Oh, and P. de Groen. Quadrant coverage histogram: a new method for measuring quality of colonoscopic procedures. In *Proc. 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3470–3473, 2007.

[67] G. N. Khan and D. F. Gillies. Detecting insertion direction for an automatic endoscope. In B. Barber, D. Cao, D. Qin, and G. Wagner, editors, *MEDINFO 89*, pages 455–459. North-Holland, 1989.

[68] G. N. Khan. *Machine Vision for Endoscope Control and Navigation.* PhD thesis, Imperial College, London, 1989.

[69] L. E. Sucar and D. F. Gillies. Knowledge-based assistant for colonscopy. In *Proc. 3rd International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pages 665–672, New York, NY, USA, 1990. ACM.

[70] L. E. Sucar, D. F. Gillies, and H. U. Rashid. Integrating shape from shading in a gradient histogram and its application to endoscope navigation. In *Proc. 5th International Conference on Artificial Intelligence*, 1992.

[71] H. U. Rashid and P. Burger. Differential algorithm for the determination of shape from shading using a point light source. *Image and Vision Computing*, 10(2):119 – 127, 1992.

[72] C. K. Kwoh and D. F. Gillies. Using fourier information for the detection of the lumen in endoscope images. In *Proc. IEEE Region 10's Ninth Annual International Conference. Theme: Frontiers of Computer Technology*, pages 981–985, Aug 1994.

[73] C. K. Kwoh, G. N. Khan, and D. F. Gillies. Automated endoscopic navigation and advisory system from medical imaging. In *Medical Imaging 1999: Physiology and Function from Multidimensional Images*, volume 3660-1, pages 214–224. SPIE, 1999.

[74] S. Krishnan, C. Tan, and K. Chan. Closed-boundary extraction of large intestinal lumen. In *Proc. 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 610–611 vol.1, Nov 1994.

[75] D.-M. Tsai and Y.-H. Chen. A fast histogram-clustering approach for multi-level thresholding. *Pattern Recognition Letters*, 13(4):245 – 252, 1992.

[76] S. Kumar, K. Asari, and D. Radhakrishnan. A new technique for the segmentation of lumen from endoscopic images by differential region growing. In *Proc. 42nd Midwest Symposium on Circuits and Systems*, volume 1, pages 414–417 vol. 1, 1999.

[77] M. Cheriet, J. Said, and C. Suen. A recursive thresholding technique for image segmentation. *IEEE Transactions on Image Processing*, 7(6):918 –921, Jun 1998.

[78] S. Kumar, K. Asari, and D. Radhakrishnan. Real-time automatic extraction of lumen region and boundary from endoscopic images. *Medical and Biological Engineering and Computing*, 37(1):600–604, 1999.

[79] S. Kumar, K. V. Asari, and D. Radhakrishnan. Online extraction of lumen region and boundary from endoscopic images using a quad structure. In *Proc. 7th International Conference on Image Processing and Its Applications*, volume 2, pages 818–822, July 13–15, 1999.

[80] K. V. Asari. A fast and accurate segmentation technique for the extraction of gastrointestinal lumen from endoscopic images. *Medical Engineering & Physics*, 22(2):89 – 96, 2000.

[81] H. Tian, T. Srikanthan, and K. Asari. A recursive otsu-iris filter technique for high-speed detection of lumen region from endoscopic images. In *Proc. 30th Applied Imagery Pattern Recognition Workshop*, pages 182–186, Oct 2001.

[82] H. Tian, T. Srikanthan, and K. Vijayan Asari. Automatic segmentation algorithm for the extraction of lumen region and boundary from endoscopic images. *Medical and Biological Engineering and Computing*, 39(1):8–14, 2001.

[83] M. P. Tjoa, S. M. Krishnan, C. J. Yap, S. Swaminathan, and P. Wang. Application of artificial neural networks for classification of colonoscopic images. In *Proc. Asia-Pacific Conference on Circuits and Systems*, volume 2, pages 227–230, October 28–31, 2002.

[84] C. Lim, H. Tian, and T. Srikanthan. An automated technique for high speed segmentation of endoscopic images. In *Proc. International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 186–189, 2004.

[85] H. Chettaoui, G. Thomann, C. B. Amar, and T. Redarce. Extracting and tracking colons "pattern" from colonoscopic images. In *Proc. 3rd Canadian Conference on Computer and Robot Vision*, page 65, June 07–09, 2006.

[86] S. Krishnan, X. Yang, K. L. Chan, and P. Goh. Region labeling of colonoscopic images using fuzzy logic. In *Proc. 1st Joint Conference of the IEEE Engineering in Medicine and Biology Society and the Biomedical Engineering Society*, volume 2, pages 1149 vol.2–, Oct 1999.

[87] R. Schettini. A segmentation algorithm for color images. *Pattern Recognition Letters*, 14(6):499–506, 1993.

[88] Y. Cao, D. Liu, W. Tavanapong, J. Wong, J. Oh, and P. de Groen. Automatic Classification of Images with Appendiceal Orifice in Colonoscopy Videos. In *Proc. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2349 – 2352, 2006.

[89] X. Liu, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Automated measurement of quality of mucosa inspection for colonoscopy. *Procedia Computer Science*, 1(1):951 – 960, 2010. ICCS 2010.

[90] X. Zabulis, A. Argyros, and D. Tsakiris. Lumen detection for capsule endoscopy. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3921–3926, 2008.

[91] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790 –799, Aug 1995.

[92] H. Messmann and J. Barnert. *Atlas of Colonoscopy: Techniques, Diagnosis, Interventional Procedures.* George Thieme Verlag, 2006.

[93] S. Ameling, S. Wirth, N. Shevchenko, T. Wittenberg, D. Paulus, and C. Münzenmayer. Detection of lesions in colonoscopic images: A review. In *Proc. World Congress on Medical Physics and Biomedical Engineering*, volume 254 of *IFMBE Proceedings*, pages 995–998. Springer Berlin Heidelberg, 2010.

[94] J. Bernal, F. Vilariño, and J. Sánchez. Towards intelligent systems for colonoscopy. In P. Miskovitz, editor, *Colonoscopy.* InTech, 2011.

[95] J. A. DiPalma. Preparation for colonoscopy. In J. D. Wayne, D. K. Rex, and C. B. Wiliams, editors, *Colonoscopy: Principles and Practice*, chapter 18, pages 210–219. Wiley-Blackwell, 2003.

[96] S. D. Wexner, T. Force, D. E. Beck, T. H. Baron, R. D. Fanelli, N. Hyman, B. Shen, and K. E. Wasco. A consensus document on bowel preparation before colonoscopy: Prepared by a task force from the american society of colon and rectal surgeons (ASCRS), the american society for gastrointestinal endoscopy (ASGE), and the society of american gastrointestinal and endoscopic surgeons (SAGES). *Gastrointestinal Endoscopy*, 63(7):894 – 909, 2006.

[97] J. Belsey, O. Epstein, and D. Heresbach. Systematic review: oral bowel preparation for colonoscopy. *Alimentary Pharmacology and Therapeutics*, 25(4):373, 2007.

[98] C. A. Aronchick, W. H. Lipshutz, S. H. Wright, F. Dufrayne, and G. Bergman. A novel tableted purgative for colonoscopic preparation: Efficacy and safety comparisons with colyte and fleet phospho-soda. *Gastrointestinal Endoscopy*, 52(3):346 – 352, 2000.

[99] A. Rostom and E. Jolicoeur. Validation of a new scale for the assessment of bowel preparation quality. *Gastrointestinal Endoscopy*, 59(4):482 – 486, 2004.

[100] A. H. Calderwood and B. C. Jacobson. Comprehensive validation of the boston bowel preparation scale. *Gastrointestinal Endoscopy*, 72(4):686 – 692, 2010.

[101] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen. Stool detection in colonoscopy videos. In *Proc. 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3004–3007, Aug. 2008.

[102] R. Duda, P. Hart, and D. Stork. *Pattern classification.* Wiley New York, 2001.

[103] F. Vilariño, P. Spyridonos, O. Pujol, J. Vitrià, and P. Radeva. Automatic detection of intestinal juices in wireless capsule video endoscopy. In *Proc. 18th International Conference on Pattern Recognition*, volume 4, pages 719–722, 2006.

[104] B. Jähne. *Digital Image Processing, 4th edition.* Springer, Berlin, 1997.

[105] D. R. Uecker, C. Lee, Y. Wang, and Y. Wang. Automated instrument tracking in robotically assisted laparoscopic surgery. *Journal of Image Guided Surgery*, 1(6):308–325, 1995.

[106] S. Voros, J. Long, and P. Cinquin. Automatic Detection of Instruments in Laparoscopic Images: A First Step Towards High-level Command of Robotic Endoscopic Holders. *The International Journal of Robotics Research*, 26(11-12):1173, 2007.

[107] Y.-F. Wang, D. R. Uecker, and Y. Wang. A new framework for vision-enabled and robotically assisted minimally invasive surgery. *Computerized Medical Imaging and Graphics*, 22(6):429 – 437, 1998.

[108] Y. Cao, D. Li, W. Tavanapong, J. Oh, J. Wong, and P. de Groen. Parsing and browsing tools for colonoscopy videos. In *Proc. 12th annual ACM international conference on Multimedia*, pages 844–851. ACM New York, NY, USA, 2004.

[109] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision.* Thomson Learning, 3rd, international student edition, 2008.

[110] Y. Cao, D. Liu, W. Tavanapong, J. Wong, J. Oh, and P. de Groen. Computer-aided detection of diagnostic and therapeutic operations in colonoscopy videos. *IEEE Transactions on Biomedical Engineering*, 54(7):1268–1279, 2007.

[111] M. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962.

[112] R. Summers, J. Yao, and C. Johnson. CT colonography with computer-aided detection: Automated recognition of ileocecal valve to reduce number of false-positive detections. *Radiology*, 233(1):266, 2004.

[113] L. Lu, A. Barbu, M. Wolf, J. Liang, L. Bogoni, M. Salganicoff, and D. Comaniciu. Simultaneous detection and registration for ileo-cecal valve detection in 3D CT colonography. In *Proc. 10th European Conference on Computer Vision, Springer LNCS 5305*, pages 465–478, 2008.

[114] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Edge cross-section features for detection of appendiceal orifice appearance in colonoscopy videos. In *Proc. 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3000–3003, Aug. 2008.

[115] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum. Detection of blood vessels in retinal images using two-dimensional matched filters. *IEEE Transactions on Medical Imaging*, 8(3):263–269, 1989.

[116] V. Mahadevan, H. Narasimha-Iyer, B. Roysam, and H. Tanenbaum. Robust model-based vasculature detection in noisy biomedical images. *IEEE Transactions on Information Technology in Biomedicine*, 8(3):360–376, 2004.

[117] Y. Wang, W. Tavanapong, J. S. Wong, J. Oh, and P. C. de Groen. Detection of quality visualization of appendiceal orifices using local edge cross-section profile features and near pause detection. *IEEE Transactions on Biomedical Engineering*, 57(3):685 –695, march 2010.

[118] J. B. Marshall and D. N. Brown. Photodocumentation of total colonoscopy: how successful are endoscopists? do reviewers agree? *Gastrointestinal Endoscopy*, 44(3):243 – 248, 1996.

[119] D. K. Rex. Still photography versus videotaping for documentation of cecal intubation: a prospective study. *Gastrointestinal Endoscopy*, 51(4):451 – 459, 2000.

[120] Y. Cao, W. Tavanapong, K. Kim, J. Wong, J. Oh, and P. de Groen. A framework for parsing colonoscopy videos for semantic units. In *Proc. IEEE International Conference on Multimedia and Expo*, volume 3, pages 1879–1882 Vol.3, June 2004.

[121] J. Oh, S. Hwang, Y. Cao, W. Tavanapong, D. Liu, J. Wong, and P. de Groen. Measuring objective quality of colonoscopy. *IEEE Transactions on Biomedical Engineering*, 56(9):2190–2196, Sept. 2009.

[122] J. Oh, M. A. Rajbal, J. K. Muthukudage, W. Tavanapong, J. Wong, and P. C. de Groen. Real-time phase boundary detection in colonoscopy videos. In *Proc. 6th International Symposium on Image and Signal Processing and Analysis*, pages 724–729, September 16–18, 2009.

[123] S. R. Stanek, W. Tavanapong, J. S. Wong, J. Oh, and P. C. de Groen. Automatic real-time capture and segmentation of endoscopy video. In *Proc. Medical Imaging 2008: PACS and Imaging Informatics*, volume 6919-1, page 69190X. SPIE, 2008.

[124] D. K. Iakovidis, S. Tsevas, D. Maroulis, and A. Polydorou. Unsupervised summarisation of capsule endoscopy video. In *Proc. 4th IEEE International Conference on Intelligent Systems*, volume 1, pages 3–15–3–20, September 6–8, 2008.

[125] S. Tsevas, D. Iakovidis, D. Maroulis, and E. Pavlakis. Automatic frame reduction of wireless capsule endoscopy video. In *Proc. 8th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–6, Oct. 2008.

[126] O. Okun and H. Priisalu. Unsupervised data reduction. *Signal Processing*, 87(9):2260 – 2267, 2007.

[127] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[128] D. D. Lee and H. S. Seung. Unsupervised learning by convex and conic coding. *Advances in Neural Information Processing Systems*, 9:515–521, 1997.

[129] M. Mackiewicz, J. Berens, and M. Fisher. Wireless capsule endoscopy color video segmentation. *IEEE Transactions on Medical Imaging*, 27(12):1769 –1781, 2008.

[130] J. Cunha, M. Coimbra, P. Campos, and J. Soares. Automated topographic segmentation and transit time estimation in endoscopic capsule exams. *IEEE Transactions on Medical Imaging*, 27(1):19–27, 2008.

[131] F. Vilariño, S. Ameling, G. Lacey, A. Ghosh, S. Patchett, and H. Mulcahy. Eye tracking search patterns in expert and trainee colonoscopists: A novel method of assessing endoscopic competency? *Gastrointestinal Endoscopy*, 69(5):AB370–AB370, 2009.

[132] W. Ahn, W. Kim, H. Woo, K. Lee, J. Cho, D. Lee, and S. Yi. Colonoscopy simulator with enhanced haptic realism and visual feedback. In *World Congress on Medical Physics and Biomedical Engineering 2006*, IFMBE Proceedings, pages 3820–3823. Springer Berlin Heidelberg, 2007.

[133] S. Buzink, A. Koch, J. Heemskerk, S. Botden, R. Goossens, H. de Ridder, E. Schoon, and J. Jakimowicz. Acquiring basic endoscopy skills by training on the GI Mentor II. *Surgical Endoscopy*, 21(11):1996–2003, November 2007.

[134] CSIRO Biomedical Imaging Team. Developing a next generation colonoscopy simulator. `http://aehrc.com/files/projects/colonoscopy/aehrc_whitepaper_colonoscopy_v1.pdf`: [Last accessed: 24 Sep 2012].

[135] A. Ferlitsch, P. Glauninger, A. Gupper, M. Schillinger, M. Haefner, A. Gangl, and R. Schoefl. Evaluation of a virtual endoscopy simulator for training in gastrointestinal endoscopy. *Endoscopy*, 34(9):698–702, 2002.

[136] E. M. Ritter, D. A. McClusky, A. B. Lederman, A. G. Gallagher, and C. D. Smith. Objective psychomotor skills assessment of experienced and novice flexible endoscopists with a virtual reality simulator. *Journal of Gastrointestinal Surgery*, 7(7):871 – 878, 2003.

[137] S. Adamsen, P. M. Funch-Jensen, A. M. Drewes, J. Rosenberg, and T. P. Grantcharov. A comparative study of skills in virtual laparoscopy and endoscopy. *Surgical Endoscopy*, 19(2):229–234, February 2005.

[138] T. Doyle. Interlaced to sequential conversion for EDTV applications. In *Proc. 2nd Int. Workshop Signal Processing of HDTV*, pages 412–430, 1988.

[139] K. C. Huh and D. K. Rex. Missed neoplasms and optimal colonoscopic withdrawal technique. In J. D. Waye, D. K. Rex, and C. B. Williams, editors, *Colonoscopy: Principles and Practice*, chapter 41, pages 560–571. Wiley-Blackwell, 2009.

[140] F. Vogt, D. Paulus, B. Heigl, C. Vogelgsang, H. Niemann, G. Greiner, and C. Schick. Making the invisible visible: Highlight substitution by color light fields. In *Proc. 1st European Conference on Colour in Graphics, Imaging, and Vision*, pages 352–357, 2002.

[141] R. S. of ITU. *Recommendation ITU-R BT.601-7: Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios.* International Telecommunication Union, 3 2011.

[142] E. Decencière. Motion picture restoration using morphological tools. In P. Maragos, R. W. Schafer, and M. Butt, editors, *Mathematical morphology and its applications to image and signal processing*, pages 361–368. Kluwer Academic Publishers, 1996.

[143] O. Buisson, B. Besserer, S. Boukir, and F. Helt. Deterioration detection for digital film restoration. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 78 –84, 1997.

[144] J. Han and M. Kamber. *Data mining: concepts and techniques.* Morgan Kaufmann, 2006.

[145] T. K. Shih and R.-C. Chang. Digital inpainting - survey and multilayer image inpainting algorithms. In *Proc. International Conference on Information Technology and Applications*, volume 1, pages 15–24, Los Alamitos, CA, USA, 2005. IEEE Computer Society.

[146] A. Kokaram. On missing data treatment for degraded video and film archives: a survey and a new bayesian approach. *IEEE Transactions on Image Processing*, 13(3):397–415, 2004.

[147] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.

[148] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Adaptive Computation and Machine Learning. MIT Press, 2002.

[149] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[150] I. Daubechies and C. Mass. *Ten lectures on wavelets Regional conference series in applied mathematics.* Society for industrial and applied mathematics, Philadelphia, 1995.

[151] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761 – 767, 2004. British Machine Vision Computing 2002.

[152] R. Kimmel, C. Zhang, A. Bronstein, and M. Bronstein. Are MSER features really interesting? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2316 –2320, nov. 2011.

[153] A. Vedaldi. *An Implementation of Multi-Dimensional Maximally Stable Extremal Regions*, 2007.

[154] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[155] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[156] K. Krippendorff. Reliability in content analysis. *Human Communication Research*, 30(3):411–433, 2004.

[157] M. Lombard, J. Snyder-Duch, and C. C. Bracken. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604, 2002.

[158] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology.* Sage, 2004.

[159] A. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.

[160] J. Petrini. Continuous quality improvement in colonoscopy. In J. D. Waye, D. K. Rex, and C. B. Williams, editors, *Colonoscopy: Principles and Practice, 2nd Ed.*, chapter 3, pages 41–54. Wiley-Blackwell, 2009.

[161] J. G. Adair. The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, 69(2):334 – 345, 1984.

[162] K. Leonard and M. C. Masatu. Outpatient process quality evaluation and the Hawthorne effect. *Social Science & Medicine*, 63(9):2330 – 2340, 2006.

[163] D. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures.* Chapman & Hall/CRC, 2004.

[164] K. Krippendorff. Computing Krippendorff's alpha reliability. Technical report, Departmental Papers (The Annenberg School for Communication at University of Pennsylvania), 2007.

Appendix A

Statistical Analysis

A.1 Measures of Association Between Pairs of Data

The common purpose of measures of association is to measure, to what degree two variables are statistically dependent. Depending on the nature of the data and the purpose of measuring association, we have chosen a number of different measures of association. This section contains the basic definitions of the measures used throughout the thesis. For a more detailed treatment, the reader is referred to [163].

In the following, the variables to be analysed are denoted by X and Y. The equations refer to N observations X_i,Y_i of the variables, with the sample means $\overline{X}$ and $\overline{Y}$.

A.1.1 Levels of Measurement

Data representations can be categorised into 4 different levels of measurement [163]: *nominal* (or *categorical*) data, *ordinal* data, *interval* data and *ratio* data. Nominal and ordinal data represent a qualitative measurement, while interval and ratio data result from quantitative measurement.

At the nominal level, the data represents categories without any defined ordering. Examples are different currencies or musical genres. At the ordinal level, measurements have a defined rank ordering, in the sense that one measurement is higher, lower or equal to another, while the distance between adjacent ranks is unknown. A common example is the rating of an item using the scale 1-poor, 2-fair, 3-good and 4-excellent. Continuous data can also be ordinal, if we can not assume a linear relationship between the measurement scale and the measured quantity.

At the interval level, equal distances on the measurement scale correspond to equal distances of the measured quantity. Interval level scales have no meaningful zero point. An example is the celsius temparature scale, where the zero point was initially set arbitrarily at the freezing point of water. Consequently, such measures have no meaningful notion of ratios (20°C are not twice as warm as 10°C). Is a meaningful zero point defined, such as for the metric scale for length, the data is called ratio data.

A.1.2 Pearson Product-Moment Correlation Coefficient

The Pearson product moment correlation coefficient r is a measure of linear dependence between two variables. For a sample, it is computed as the sample covariance divided by the product of sample variances:

$$r = \frac{\sum_{i=1}^{N}(X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2} \cdot \sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}}. \tag{A.1}$$

r is meaningful only for interval and ratio data due to the use of variance and covariance. It is, furthermore, based on the assumption that the sample was drawn randomly from a bivariate normal distribution. r is often reported regardless of violation of these assumtions. We are dealing mostly with ordinal data throughout the thesis. We are reporting r therefore only for comparison due to its common use in similar applications.

A.1.3 Kendall's τ Rank Correlation Coefficient

Kendall's τ is a non-parametric measure of association for ordinal data. For a sample, its value is computed from the pairs of examples $p_{i,j} = \{\{X_i, Y_i\}, \{X_j, Y_j\}\}, i, j \in \{1..N\}$ as the difference between the number of concordant pairs C and the number of discordant pairs D, divided by the total number of pairs:

$$\tau = \frac{C - D}{0.5N(N-1)}. \tag{A.2}$$

A pair of examples $p_{i,j}$ is...:

- *concordant*, if $[(X_i > Y_i) \wedge (X_j > Y_j)] \vee [(X_i < Y_i) \wedge (X_j < Y_j)]$.
- *discordant*, if $[(X_i > Y_i) \wedge (X_j < Y_j)] \vee [(X_i < Y_i) \wedge (X_j > Y_j)]$.
- *tied*, if $(X_i = X_j) \vee (Y_i = Y_j)$.

Ties occur frequently in our analysis. We therefore use τ_B, a variant of τ that corrects for ties. It is defined as

$$\tau_B = \frac{C - D}{\sqrt{(P - T_1)(P - T_2)}} \tag{A.3}$$

with

$$
\begin{aligned}
P &= 0.5N(N-1) \\
T_1 &= 0.5\sum_i t_i(t_i-1) \\
T_2 &= 0.5\sum_j u_j(u_j-1)
\end{aligned}
$$

For the computation of t_i and u_j, ties are counted separately for each variable and arranged into groups. Each group of ties comprises all pairs of examples for which the variable under consideration has the same value (regardless of the value of the other variable). Then, t_i is the number of tied values in the i^{th} group of ties for the first variable, and u_j is the number of tied values in the j^{th} group of ties for the second variable.

A.2 Krippendorff's α: A Measure of Inter-Rater Reliability

Krippendorff's α is a universally applicable, chance-corrected measure of inter-rater reliability. It combines a number of reliability measures, each of which applying to data at a specific level of measurement, and makes adjustments for better comparability between different types of data. α adjusts itself to varying sample sizes and can handle missing data. The only prerequisite for computing α is that the data be reliability data.

Reliability data "result from duplicating the process of describing, categorizing, or measuring a sample of data obtained from the population of data whose reliability is in question." [156]. Duplication can be the assessment of the sample by jointly instructed, but independently working assessors.

In the case of the analysis in Chapter 6, duplication is provided by the 2 assessors who both participated in a training course on the subject and were jointly instructed on the particular experiment. Krippendorff's α is, however, applicable to any number of duplicate assessments.

In its most general form, α is defined as [164]

$$\alpha = 1 - \frac{D_o}{D_e}, \tag{A.4}$$

with D_o and D_e being the observed and expected disagreement, respectively. It can take on values between -1 and 1. $\alpha = 1$ means that there is perfect agreement, $\alpha = 0$ that agreement is as good as chance agreement, and $\alpha < 0$ that there is systematic disagreement. The computation of D_o and D_e is as follows:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck} \delta^2_{ck}, \tag{A.5}$$

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c n_k \delta^2_{ck}. \tag{A.6}$$

To obtain n, o_{ck}, n_c, n_k and δ^2_{ck}, the coincidence matrix of the set of reliability data needs to be computed. A coincidence matrix is a tabulation of the frequency of occurrence of all pairs of assessments in the reliability data. Consider the following example of ratings on an ordinal scale by two assessors (A_1, A_2):

Unit	1	2	3	4	5	6	7	8	9	10
A_1	1	4	2	2	1	3	4	3	1	4
A_2	1	4	3	2	2	3	3	2	3	3

The entries o_{ck} in the coincidence matrix are obtained by counting, how many pairs of values {c,k} (c, $k \in \{1, 2, 3, 4\}$ in our example) exist in the data. For two assessors, each unit is entered twice into the matrix, e.g., unit 3 counts both as a {3,2}-pair and as a {2,3}-pair, and unit 1 counts as two {1,1} pairs. The values are entered into the matrix according to the following template:

	1	**2**	**3**	**4**	$\sum$
1	o_{11}	o_{12}	o_{13}	o_{14}	n_1
2	o_{21}	o_{22}	o_{23}	o_{24}	n_2
3	o_{31}	o_{32}	o_{33}	o_{34}	n_3
4	o_{41}	o_{42}	o_{43}	o_{44}	n_4
$\sum$	n_1	n_2	n_3	n_4	

For our example this results in

	1	**2**	**3**	**4**	$\sum$
1	2	1	1	0	4
2	1	2	2	0	5
3	1	2	2	2	7
4	0	0	2	2	4
$\sum$	4	5	7	4	

The δ^2_{ck} in (A.5,A.6) are coefficients arising from the used distance metric. This metric depends on the level of measurement of the data. For nominal data, for example, $\delta^2_{ck} = 0$ if $c = k$ and 1 otherwise. In our analysis in Chapter 6, we deal with ordinal data, in which case δ^2_{ck} amounts to the squared difference between the ranks of the current values, or, using quantities from the coincidence matrix,

$$\delta^2_{ck} = \left[\left(\sum_{i=\min(\{c,k\})}^{\max(\{c,k\})} n_i \right) - \frac{n_c + n_k}{2} \right]^2 .$$

We omit further details on distance metrics for other types of data, as our analysis is restricted to ordinal data. For a more detailed treatment of Krippendorff's α and the additional distance metrics, see [164] or [158].

Appendix B

JAG DOPS Assessment Form and Grade Descriptors

The following pages show the JAG DOPS assessment form and the associated grade descriptors. The forms and additional information can be found on the website of the Joint Advisory Group on GI Endoscopy (`http://www.thejag.org.uk/`).

JAG
Joint Advisory Group
on GI Endoscopy

DOPS Assessment Form

Certification of Screening Colonoscopists

Candidate

Assessor

Assessment Centre

Date (DD/MM/YYYY)

Case Number

Scale and Criteria Key

4	Highly skilled performance
3	Competent and safe throughout procedure, no uncorrected errors
2	Some standards not yet met, aspects to be improved, some errors uncorrected
1	Accepted standards not yet met, frequent errors uncorrected
n/a	Not applicable

■ Major Criteria   □ Minor Criteria

Headline Criteria Full Criteria outlined in Grade Descriptors	**Score**	**Comments**
Assessment, consent, communication		
■ Obtains informed consent using a structured approach ○ Satisfactory procedural information ○ Risk and complications explained ○ Co-morbidity ○ Sedation ○ Opportunity for questions		
■ Demonstrates respect for patient's views and dignity during the procedure		
■ Communicates clearly with patient, including outcome of procedure with appropriate management and follow up plan.		
Safety and sedation		
■ Safe and secure IV access		
■ Gives appropriate dose of analgesia and sedation and ensures adequate oxygenation and monitoring of patient		
■ Demonstrates good communication with the nursing staff, including dosages and vital signs		
Endoscopic skills during insertion and procedure		
□ Checks endoscope function before intubation		
□ Performs PR		
■ Maintains luminal view / inserts in luminal direction		
■ Demonstrates awareness of patient's consciousness and pain during the procedure and takes appropriate action		
□ Uses torque steering and control knobs appropriately		
□ Uses distension, suction and lens washing appropriately		
■ Recognises and logically resolves loop formation		
□ Uses position change and abdominal pressure to aid luminal views		
□ Completes procedure in reasonable time		
Diagnostic and therapeutic ability		
■ Adequate mucosal visualisation		
■ Recognises caecal landmarks or incomplete examination		
■ Accurate identification and management of pathology		
■ Uses diathermy and therapeutic techniques appropriately and safely		
■ Recognises and manages complications appropriately		

Case Difficulty

Extremely easy	Fairly easy	Average	Fairly difficult	Very challenging
1	2 3		4	5

JAG

Joint Advisory Group
on GI Endoscopy

Assessor Declaration

Certification of Screening Colonoscopists

To become an a Certified screening colonoscopist, the candidate must finish the two cases having achieved the following major and minor criteria

DOPS STANDARDS

MAJOR DOMAINS (14 DOMAINS)

□ We declare that the candidate received a Grade 3 or Grade 4 on all 14 major domains

□ We declare that there are **no** Grade 1 or Grade 2 scores in any of the 14 major domains.

MINOR DOMAINS (6 DOMAINS)

□ We declare that the candidate has not exceeded four grade 2's when summated across four cases.

□ We declare that there are **no** Grade 1 scores in any of the six minor domains.

CONFIDENTIAL - EXPERT GLOBAL EVALUATION

In order to help with setting standards and validating the process, please give your expert global assessment **independent** of the above grading – in other words, do you personally judge that the colonoscopist is ready to become an independent colonoscopist.

Please check one of the two boxes below.

□ The candidate should be certified for screening colonoscopy

□ The candidate should not yet be certified for screening colonoscopy

ASSESSOR SIGN OFF

We certify that ____GMC no____

□ Meets the DOPS criteria outlined on page one

□ Meets the minimum DOPS standards above

Assessor 1 **GMC number**

Assessor 2 **GMC number**

JAG

Joint Advisory Group
on GI Endoscopy

DOPS Grade Descriptors

Certification of Screening Colonoscopists

Descriptors for each grade in all four domains are given below to improve consistency of grading. The key descriptor level is Grade 3. Grade 4 assumes achievement of all components at Grade 3 level and some achievement above this.

The descriptors set expectations for the performance in each domain, but should be used as a guide – colonoscopists do not have to meet all criteria in each descriptor to achieve a grade in that domain.

ASSESSMENT, CONSENT AND COMMUNICATION

GRADE 4

Complete and full explanation in clear terms including proportionate risks and consequences with no omissions of significance, and not unnecessarily raising concerns. No jargon. Encourages questions by verbal and non verbal skills and is thoroughly respectful of individual’s views, concerns, and perceptions. Good rapport with patient. Seeks to ensure procedure is carried out with as much dignity and privacy as possible. Clear and appropriate communication throughout procedure and afterwards a thorough explanation of results and management plan.

GRADE 3

Good clear explanation with few significant omissions, covering key aspects of the procedure and complications with some quantification of risk. Little jargon, and gives sufficient opportunity for questions. Responds to individual’s perspective. Aware of and acts to maintain individual’s dignity. Appropriate communication during procedure including warning patient of probable discomfort. Satisfactory discussion of results and management plan with adequate detail.

GRADE 2

Explains procedure but with several omissions, some of significance. Little or no quantification of risk, or raises occasional unnecessary concerns. Some jargon and limited opportunity for questions or sub-optimal responses. Incomplete acknowledgement of individual’s views and perceptions. A few lapses of dignity only partially or tardily remedied. Occasional communication during the procedure and intermittent warnings of impending discomfort. Barely adequate explanation with some aspects unclear, inaccurate or lacking in detail.

GRADE 1

Incomplete explanation with several significant omissions and inadequate discussion, lacking quantification of risks or raising significant fears. Uses a lot of jargon or technical language; minimal or no opportunity for questions. Fails to acknowledge or respect individual’s views or concerns. Procedure lacks dignity and there is minimal or no communication during it. Explanation of results and management is unclear, inaccurate or lacking in detail without opportunity for discussion.

SAFETY AND SEDATION

GRADE 4

Safe and secure IV access with doses of analgesia and sedation according to patient's age and physiological state, clearly checked and confirmed with nursing staff. Patient very comfortable throughout. Oxygenation and vital signs monitored continually as appropriate, remaining satisfactory throughout or rapid and appropriate action taken if sub-optimal. Clear, relevant and proactive communication with endoscopy staff.

GRADE 3

Secure IV access with a standard cannula and appropriate dose of analgesia and sedation within current guidelines, checked and confirmed with nursing staff. Patient reasonably comfortable throughout, some tolerable discomfort may be present. Oxygenation and vital signs regularly monitored and satisfactory throughout, or appropriate action taken. Clear communication with endoscopy staff.

GRADE 2

IV access acceptable with just satisfactory analgesia and sedation incompletely confirmed or checked with nursing staff, patient too sedated or too aware and in discomfort. Oxygenation and vital signs monitored but less frequently than appropriate or parameters occasionally unsatisfactory with action taken only after prompting or delay. Intermittent or sub optimal communication with endoscopy staff.

GRADE 1

Insecure or absent IV access or butterfly used; inadequate or inaccurate check of analgesia and sedation. Patient significantly under- or over-sedated or needing use of a reversal agent because of inappropriate dosaging. Patient in discomfort much of the time, or significant periods of severe discomfort. Oxygenation and vital signs rarely or inadequately monitored and mostly ignored even if unsatisfactory. Minimal or significantly flawed communication with endoscopy staff.

ENDOSCOPIC SKILLS DURING INSERTION AND WITHDRAWAL

GRADE 4

Excellent luminal views throughout the vast majority of the examination, with judicious use of "slide-by". Skilled torque steering and well judged use of distension, suction and lens clearing. Rapid recognition and resolution of loops. Quick to use position change or other manoeuvres when appropriate. Immediately aware of patient discomfort with rapid response. Smooth scope manipulation using angulation control knobs and torque steering.

GRADE 3

Check scope functions, performs PR. Clear luminal view most of the time or uses slide-by appropriately. Appropriate use of the angulation control knobs. Uses torque steering adequately. Aids progress using distension, suction and lens washing. Recognises most loops quickly and attempts logical resolution. Good use of position changes to negotiate difficulties. Aware of any discomfort to patient and responds with appropriate actions. Timely completion of procedure, not too quickly or too slowly for the circumstances.

GRADE 2

Omits scope check or PR. Luminal views lost a little more than desirable or uses slide-by a little too long or frequently. Could torque steer usefully more often or more effectively. Some under or over distension or lack of lens washing. Recognises most loops with reasonable attempts at resolution. Use of position change or other manoeuvres occasionally late or inappropriately. Aware of and responsive to patient but may be slow to do so. Procedure slightly too fast or too slow.

GRADE 1

Omits to check scope or rectal examination. Luminal views frequently lost for long periods and pushes on regardless. Little or no use of torque steering. Under- or over-distension of bowel, or fails to attempt lens clearing. Recognises loops late or not at all and little or no structured attempt to resolve them. Inappropriate or no use of position change or other manoeuvres. Barely aware of patient’s status, or very tardy / inappropriate / no response to discomfort. Completes examination too quickly or takes far too long.

DIAGNOSTIC AND THERAPEUTIC ABILITY

GRADE 4

Excellent mucosal views throughout the majority of the procedure. Recognition of all caecal landmarks present or rapidly identifies incomplete examination. Faecal pools fully suctioned. Retroflexes in rectum. Thorough assessment and accurate identification of pathology present. Skilled and competent management of diathermy and therapeutic techniques. Rapid recognition and appropriate management of complications.

GRADE 3

Adequate mucosal visualisation with only occasional loss or sub-optimal views unless outwith control of endoscopist (e.g. stool, severe diverticular disease). Faecal pools adequately suctioned. Attempts to retroflex in rectum. Correctly identifies caecal landmarks or incomplete examination. Accurately identifies pathology and manages appropriately according to current guidelines. Correct and safe use of diathermy and therapeutic techniques. Rapid recognition of complications with safe management.

GRADE 2

Mucosal views intermittently lost for more than desirable periods. Recognises most caecal landmarks present or eventually identifies an incomplete examination. Most pathology identified with occasional missed or mis-identified lesions. Just acceptable use of diathermy and therapeutic tools with some sub optimal use. Delayed or incomplete recognition of complications or sub-optimal management.

GRADE 1

Frequent or prolonged loss of mucosal views. Incorrect identification of caecal landmarks, or fails to recognise incomplete examination. Misses significant pathology, or inappropriate management that may endanger patient or contravenes guidelines. Unsafe use of diathermy and therapeutic techniques. Fails to recognise or significantly mis-manages complications to the detriment of the patient.

Appendix C

Instruction Sheets

PARTICIPANT INSTRUCTIONS

INSTRUCTIONS FOR PERFORMERS

EXPERIMENT OUTLINE:

In this study you are asked to perform a screening colonoscopy on a simulator. Your task is to detect red markers in the colon model. Each marker consists of a unique letter surrounded by a circle. You are requested to say out loud the letter, whenever you find a new marker.

During the experiment you will first be requested to put on a light head mounted eye tracker, as well as a microphone. We will, furthermore, record you during the procedure with an external camera. Your face will not be in the picture and we ask you to wear a gown to make sure you remain anonymous.

As a first step the head mounted unit will be physically calibrated such that you feel comfortable wearing it. Next you will be shown a few calibration images with red crosses. You will be requested to follow the red crosses with your eyes while the system is being calibrated. Once these set up procedures are completed, you will be able to familiarise yourself with the colon model for 2 minutes, before you start the actual procedure.

You are encouraged to use lubricant to facilitate insertion and you have access to an assistant to administer abdominal pressure or turn the model into different positions. The experiment will end when you have completed the screening or choose to end the experiment for any other reason.

Your performance will be assessed by experts using an excerpt from a DOPS assessment form. Your identity will not be disclosed to the assessors and they will not hear you speaking.

PARTICIPANT INSTRUCTIONS

INSTRUCTIONS FOR ASSESSORS

EXPERIMENT OUTLINE:

In the first phase of this study we asked endosopists of various experience levels to perform a screening colonoscopy on a simulator. Their task was to detect red markers in the colon model. The procedures were recorded using a number of cameras, a motion sensor, and an eye-tracking device. Furthermore, information about their experience was collected together with polyp detection and cecal intubation rates, where available.

In this second phase we ask you to assess the performance of the participating endoscopists. You will watch videos of the procedures, showing both the view through the endoscope and an external view of the endoscopist performing the procedure. The assessment is done afterwards by filling out an assessment form, mostly containing criteria from the JAG DOPS (Direct Observation of Procedure and Skill) assessment form.

In preparation for the actual assessment, there will be a consensus meeting with all the assessors, where example videos will be shown and all rating criteria will be discussed in detail. This is to achieve high inter-rater reliability.

You will then get a numbered set of videos for assessment. There is no time schedule for assessment, but you should assess the videos in the order provided. We also suggest rating sessions of no longer than 60 minutes at a time in an environment where you are unlikely to be disturbed.

Appendix D

Procedure Assessment Form

COLONOSCOPY PERFORMANCE ASSESSMENT FORM

Date: □ Time: □ Video Number: □

For each question, please tick the box for the grade you consider appropriate. Please rate according to the description given for each grade. Use the comment box if the description does not include the reason why you chose the grade and also if you have any other comments. If you want to mention any special events, please add the approximate time in the video.

A) JAG DOPS

1. MAINTAINS LUMINAL VIEW / INSERTS IN LUMINAL DIRECTION

□	**4**	Excellent luminal views through the vast majority of the examination, with judicious use of "slide-by"
□	**3**	Clear luminal view most of the time or uses slide-by appropriately
□	**2**	Luminal views lost a little more than desirable or uses slide-by a little too long or frequently
□	**1**	Luminal views frequently lost for long periods and pushed on regardless

Comments:

2. USES TORQUE STEERING AND CONTROL KNOBS APPROPRIATELY

□	**4**	Skilled torque steering. Smooth scope manipulation using angulation control knobs and torque steering.
□	**3**	Appropriate use of angulation control knobs. Uses torque steering adequately.
□	**2**	Could torque steer usefully more often or more effectively.
□	**1**	Little or no use of torque steering.

Comments:

3. Recognises and logically resolves loop formation

□ **4** Rapid recognition and resolution of loops.

□ **3** Recognises most loops quickly and attempts logical resolution.

□ **2** Recognises most loops with reasonable attempts to resolution.

□ **1** Recognises loops late or not at all and little or no structured attempt to resolve them.

Comments:

4. Uses position change and abdominal pressure to aid luminal views

□ **4** Quick to use position change or other manoeuvres when appropriate.

□ **3** Good use of position changes to negotiate difficulties.

□ **2** Use of position change or other manoeuvres occasionally late or inappropriately.

□ **1** Inappropriate or no use of position change or other manoeuvres.

Comments:

5. Completes procedure in reasonable time

□ **3** Timely completion of procedure, not too quickly or too slowly for the circumstances.

□ **2** Procedure slightly too fast or too slow.

□ **1** Completes examination too quickly or takes far too long.

Comments:

6. Adequate mucosal visualisation

□ **4** Excellent mucosal views throughout the majority of the procedure.

□ **3** Adequate mucosal visualisation with only occasional loss or sub-optimal views unless beyond the control of the endoscopist.

□ **2** Mucosal views intermittently lost for more than desirable periods.

□ **1** Frequent or prolonged loss of mucosal views.

Comments:

B) Summary Scores

1. Insertion phase

□ **4** Excellent insertion performance.

□ **3** Adequate insertion.

□ **2** Minor shortcomings.

□ **1** Inadequate performance / major shortcomings.

Comments:

2. WITHDRAWAL PHASE

□ **4** Excellent withdrawal / examination performance.

□ **3** Adequate withdrawal.

□ **2** Minor shortcomings.

□ **1** Inadequate performance / major shortcomings.

Comments:

3. WHOLE PROCEDURE

□ **4** Excellent performance.

□ **3** Adequate performance.

□ **2** Minor shortcomings.

□ **1** Inadequate performance / major shortcomings.

Comments: