

UNIVERSITÀ DEGLI STUDI DI CATANIA
FACOLTÀ DI SCIENZE MATEMATICHE FISICHE E NATURALI
DOTTORATO DI RICERCA IN INFORMATICA

AUTOMATIC PATTERN CLASSIFICATION AND
STEREOSCOPIC VISION IN MEDICAL IMAGING

ALESSANDRO TORRISI

A dissertation submitted to the Department of Mathematics and Computer Science and the committee on graduate studies of University of Catania, in fulfillment of the requirements for the degree of doctorate in Computer Science.

ADVISOR
Prof. Giovanni Gallo
COORDINATOR
Prof. Domenico Cantone

XXV CICLO

Contents

1	Introduction	3
1.1	Dissertation organization	5
I	Automatic Classification of Frames from Wireless Capsule Endoscopy	7
2	Wireless Capsule Endoscopy	8
2.1	WCE system details	10
2.2	Benefits and risks of capsule endoscopy	14
2.3	Typical capsule endoscopy images	16
2.4	Manual annotation	17
2.5	The digestive tract	19
3	Literature review	22
3.1	Topographic segmentation	23
3.2	Event detection	26
3.2.1	Intestinal contractions	27
3.2.2	Intestinal juices	27
3.2.3	Bleeding detection	28
3.2.4	Anomaly detection	29
4	Information Theoretic Method	33
4.1	Entropy	34
4.2	Kolmogorov complexity	37
4.3	Algorithmic Information Distance	38

5	Ensemble Learning	41
5.1	AdaBoost	42
5.1.1	Real application	44
6	Experiments	49
6.1	Dataset	49
6.2	Information Theory based WCE video summarization	50
6.2.1	Feature extraction	50
6.2.2	Classification method	51
6.2.3	Experimental results	54
6.2.4	Conclusion	57
6.3	Lumen Detection in Endoscopic Images: A Boosting Classification Approach	58
6.3.1	Feature extraction	59
6.3.2	Classification method	63
6.3.3	Experimental results	66
6.3.4	Conclusion	72
6.4	Random Forests based WCE frames classification	73
6.4.1	Classification method	73
6.4.2	Experimental results	76
6.4.3	Conclusion	78
7	Conclusion and future work	80
 II Depth estimation in Bronchoscopic Intervention		83
8	Stereoscopic Vision	84
8.1	Stereoscopic system	85
8.1.1	Disparity	85
8.2	The correspondence problem	87
8.3	Epipolar geometry	91
8.3.1	Calibration	92

<i>CONTENTS</i>	iii
9 Literature Review	94
9.1 Stereoscopy in medicine	95
9.2 Augmented Reality in surgery	97
10 Experiments	99
10.1 Depth extraction from monocular bronchoscopy	103
10.1.1 Depth clues extraction	104
10.1.2 Experimental results	106
10.1.3 Discussion	107
10.2 3D reconstruction in virtual reality	110
10.2.1 Depth clues extraction	112
10.2.2 Augmented Reality	114
10.2.3 Discussion	116
10.3 Stereoscopic bronchoscope prototype	119
10.3.1 Hardware	119
10.3.2 Software	121
10.3.3 Experimental results	128
11 Conclusion and future work	132
12 Bibliography	135

List of Figures

2.1	Graphical scheme of the endocapsule.	12
2.2	The antennas array that transmits the capsule’s signal to a recorder worn by the patient.	13
2.3	A typical Wireless Capsule Endoscopy frame.	16
2.4	Different examples of capsule endoscopy scenarios.	17
2.5	Rapid Reader exam annotation software developed by Given Imaging.	18
2.6	A schematic illustration of the human GI tract.	20
2.7	The lower GI tract.	21
3.1	Examples of endoscopic frames showing intestinal juices.	28
3.2	Examples of endoscopic frames showing ulcers.	30
4.1	Grouping property of the entropy.	35
4.2	Entropy of a binary source.	36
5.1	AdaBoost pseudo-code.	43
5.2	Haar features proposed by Viola-Jones for face detection.	44
5.3	Integral image representation.	46
5.4	Evaluation of a two-rectangle feature.	46
5.5	Cascade of strong classifiers.	48
6.1	Example of an image extracted from a WCE video.	50
6.2	A schematic illustration of the <i>Textons</i> method.	52
6.3	Representation of frames as a “bag of visual words”.	53
6.4	The computation of function $Score(i)$	54

6.5	Two examples of sequences of consecutive frames.	55
6.6	Percentage of events and not-events in a WCE video.	56
6.7	Two ROC curves compare the performance of tested methods.	56
6.8	Examples of events found with the Information-theoretic method.	57
6.9	Examples of <i>lumen</i> and <i>not lumen</i> frames extracted from a WCE video.	58
6.10	The three kinds of features proposed for lumen detection.	61
6.11	Schematic representation of Haar-based features used for lu- men detection.	63
6.12	Schematic representation of the scales used for each feature.	64
6.13	Example of a cascade of strong classifiers obtained in the ex- periments.	68
6.14	ROC curve for each dataset obtained by varying the stiffness threshold of each classifier from 0.1 to 1.	70
6.15	Example of some false positives detected by the system.	70
6.16	The three features used in the proposed method.	74
6.17	Comparison of recall and precision rate as a function of the number of trees in the forest.	77
6.18	Percentage distribution of the three types of features in the final classifier.	78
6.19	Some misclassified of our classifier, false positives (a) and false negatives (b) respectively.	79
8.1	Example of a binocular stereo system.	85
8.2	Scheme of a binocular stereo system with parallel optical axes.	86
8.3	Projections of two points in a binocular stereo system.	88
8.4	An example of a standard stereo pair.	89
8.5	Epipolar geometry.	92
9.1	The surgical robot “DaVinci”.	96
9.2	Schematic representation of the VisionSense technology.	97
10.1	The three type of bronchoscopic images used in the experiments.	100
10.2	Example of application of the region growing technique.	106

10.3 Disparity maps obtained using a monocular bronchoscopic video. 108

10.4 A typical real bronchoscopic image. 110

10.5 Scheme of the canonical stereo system used in virtual reality. . 111

10.6 Examples of stereo images extracted from the proposed virtual
model. 112

10.7 Examples of depth maps estimated with the proposed method. 115

10.8 Color depth maps integrated in the reference images. 116

10.9 Hardware information of the prototype of stereo bronchoscope. 120

10.10 Hardware configuration of the prototype of stereo bronchoscope. 122

10.11 The prototype of flexible stereo bronchoscope. 123

10.12 Checkerboard pattern used for the calibration of the stereo
system. 124

10.13 Selection of the four angles of the checkerboard. 124

10.14 Extrinsic parameters of the stereo system. 125

10.15 Image rectification. 126

10.16 GUI of the acquisition software. 127

10.17 The simulation dummy used in the experiments. 129

10.18 Disparity maps obtained using the prototype of flexible stereo
bronchoscope. 130

11.1 Two examples of Augmented Reality effects that can be used
in the bronchoscopic context. 134

List of Tables

2.1	The first generation of video-capsules produced by Given Imaging.	11
6.1	Summary of experimental results.	57
6.2	Features number per scale.	65
6.3	Details on trained cascades using ten different training sets. . .	67
6.4	Classification results using Boosting.	69
6.5	Classification results using Support Vector Machine.	72
6.6	Classification Results.	78

List of Abbreviations

AdaBoost	Adaptive Boosting
CE	European Community
DCT	Discrete Cosine Transform
FDA	Food and Drug Administration
GI	GastroIntestinal
KNN	k-nearest neighbor
LBP	Local Binary Pattern
M2A	Mouth 2 Anus
MLP	Multilayer Perceptron
NCD	Normalized Compression Distance
NID	Normalized Information Distance
PAC	Probably Approximately Correct
PCA	Principal Component Analysis
RFID	Radio Frequency Identification
SBI	Suspected Blood Indicator
SVM	Support Vector Machine
WCE	Wireless Capsule Endoscopy

*Every passing minute is another
chance to turn it all around*

Chapter 1

Introduction

Medical imaging is a generic term used to define the use of medical practices to create images of the human body for clinical purpose. Today it includes a wide range of different techniques and these have greatly enhanced the quantity and quality of information available in the clinical practice. The clinician may now obtain a comprehensive view of internal structures of the human body, such as heart, kidney, lung, gut and so on. Computer assistance plays a relevant role in all these clinical applications. Each imaging technique is indeed associated with some kind of specialized workstation which maintains the appropriate tools for manipulating images, performing measurements and extracting relevant information from the available data. The major strength in the application of computers to medical imaging hence is the use of Computer Vision and Image Processing techniques to automate some specific analysis tasks.

Among the thousands of possible areas, in this dissertation we exploit the current Computer Vision technologies to propose new methods in two research fields: “Automatic Classification of Frames from Wireless Capsule Endoscopy” and “Depth Estimation in Bronchoscopic Intervention”. In both cases the exploration of tubular internal structures of the human body through the analysis of endoscopic images asks for innovative and “smart” algorithms to translate the rough image data into useful information for the doctors.

Automatic Classification of Frames from Wireless Capsule Endoscopy

Wireless Capsule Endoscopy (WCE) is a diagnostic technique used to explore intestinal regions which are difficult to reach with traditional endoscopy. The large number of images produced by this technology requires the use of computer-aided tools to select only meaningful frames to speed up the analysis time by the expert. In the first part of this dissertation a machine learning system to automatically categorize the frames in WCE videos is proposed. Our research focus in two different classification/detection tasks. In particular, we tackle the problem of the automatic detection of sudden changes in endoscopic video sequences in order to help the clinician to locate only the relevant frames for diagnostic purpose. The second problem is the automatic detection of specific events such as the intestinal contractions, that are often related to certain gastrointestinal disorders. It should be known that the interpretation of a medical examination by an expert is strongly related to his/her experience. The presence of low-skilled staff, together with the potential distractions of the observer, can affect directly the final report of the examination. Performing a computer-aided analysis or use a fully automated one as a second opinion is hence very useful. These considerations motivated the research of classification algorithms that are addressed in this dissertation.

Depth Estimation in Bronchoscopic Intervention

3D vision systems are currently used for enhancing depth perception and to provide a greater immersive experience for different research domains. Other than for entertainment, stereo viewing has being proposed for medical applications even if real applications are at the present time at their begin. In the second part of this dissertation we intend to exploit the stereo-camera setup available in stereo-viewing systems for 3D reconstruction. In particular, we aim at extracting depth information of the bronchial scene observed through an experimental stereo bronchoscopic probe. Although we focus on bronchoscopic images, our ideas can be applied to different endoscopic procedures. In particular, the information provided by a depth map representation can

be used in different ways in an endoscopic station and it might be useful to improve the visual navigation and surgical intervention. The main important advantage of the 3D reconstruction that we try to achieve is to enable the use of Augmented Reality to support the endoscope teleguide. As a proof of concept for this research, we initially report the experiments that we have conducted on a bronchoscopic video obtained from the application of a standard monocular bronchoscope. In a second section, we describe a simulation of virtual environment of a stereo-setup system over a synthetic model of a tracheo-bronchial tree. Finally, we introduce a real prototype of a stereo bronchoscope, i.e., a flexible probe equipped with a couple of micro cameras and a light source. This technological step has been developed at the labs of the “School of Engineering and technology”, University of Hertfordshire (UK), where the author of this dissertation has spent a significative time of his graduate studies, participating to this project.

1.1 Dissertation organization

This dissertation is structured in two parts: throughout the first part, which comprises the first seven chapters, we introduce the research conducted to automatically detect specific events in endoscopic images. In particular, Chapter 2 describes in detail the WCE technique that provides the ensemble of endoscopic images under examination. Chapter 3 reports and discusses the relevant works that have already been published regarding the capsule endoscopy technology. In this occasion, it is possible to know which classification problems are typically addressed and the algorithms employed to solve them. The following two chapters (Chapter 4 and Chapter 5) discuss the theory behind our experiments: Information Theory and Ensemble Learning respectively. The experiments conducted to verify our proposals are reported in Chapter 6. Chapter 7 closes the first part of this dissertation with the general discussion of our work. It also discusses which future activities may be taken to improve the results obtained insofar.

The second part of the dissertation, which includes the remaining chapters, is devoted to study the Stereoscopic Vision in the context of endoscopic imag-

ing. Chapter 8 gives to the reader some useful information on Stereoscopy. It is explained how it is possible to obtain depth clues from a pair of stereo images. It also gives some useful information on the calibration step needed to extract the camera information and to bring a pair of rough images in a standard stereo form. A literature review about Stereoscopy applied in medical devices is reported in Chapter 9. Chapter 10 reports the experiments conducted to test the validity of our proposal. Finally, Chapter 11 draws the conclusion and closes this dissertation.

Part I

Automatic Classification of Frames from Wireless Capsule Endoscopy

Chapter 2

Wireless Capsule Endoscopy

Endoscopy is the most prevailing modality for diagnosing gastrointestinal disorders. Nowadays, there are several different endoscopic procedures varying from colonoscopy and enteroscopy to full intraoperative endoscopy. The traditional “push” endoscopy involves the use of a surgical probe-tube equipped with a micro camera and a light source. The physician conducts the probe by moving it forward along the bowels and examines the recorded images projected on a screen. Modern endoscopes also contain an accessory channel, which allows to insert medical instruments to take tissue sample and perform endoscopic resections. Although this procedure is efficient both for diagnostic and therapeutic purposes, it is usually limited by the depth of the insertion of the scope allowing only the exploration of the upper small intestine. The exam also requires the presence of a qualified staff and the need of hospitalization and sedation of the patient.

While progressive size reductions in probes and imaging enhancements are enabling ongoing technical improvements in endoscopes, the rigid structure and the thickness of the probe do not allow the exhaustive exploration of long and convoluted regions like the small intestine. Limitations of current endoscopic techniques in the identification of small bowel disorders have prompted a search for alternative technologies. For this reason, a new technique called Wireless Capsule Endoscopy (WCE) [1],[2] has been introduced as a new less invasive and painless kind of endoscopy.

WCE employs the use of a pill-shaped device that is swallowed by the patient and it is propelled through the gut by the physiological intestinal peristalsis. The front-end of the capsule is equipped with a tiny camera and a transmitter wirelessly sends the recorded images to an external receiver. Once the study is finished, the recorded movie can be easily downloaded into a workstation with the appropriate software for its posterior analysis by the expert.

As the quality of the images obtained with WCE improves, this technology is strongly elective to detect abnormalities such as ulcers, bleedings or the presence of tumors in the small intestine. The exam is also less uncomfortable to the patient because it is required only to swallow a capsule. Once activated, the capsule approximately captures two frames per second; WCE operates for about 8 hours, that corresponds to the lifetime of the battery of the capsule, reaching up to 50000 useful images at the end of the examination. Images taken during the entire route of the capsule through the intestine are successively analyzed by an expert. He/She may spend up to one or more hours to gather the relevant information for a proper diagnosis. This greatly limits the use of the capsule technology as a diagnostic routine tool. Recognition of frames displaying a pathology is indeed a hard problem, and pre-processing of the whole ensemble of the frames to eliminate those that do not carry relevant information is a much needed step.

The advent of WCE endoscopic imaging technique has led to the development of a new branch of computer-aided support systems. Such systems may be deployed using Computer Vision techniques to assist a medical expert in improving the accuracy and the annotation times of medical diagnosis. Some typical tasks that can be facilitated by the use of these computer-aided tools are related to anomaly detection and categorization, data reduction and clustering, and automatic topographic segmentation of an endoscopic video.

This chapter reviews the fundamentals of WCE. Special care is paid to the structure of the capsule, with the relative benefits and drawbacks. After presenting some basic theoretical notions, we focus on the images we get out of this imaging technique, analyzing how these ones differ according to the digestive tract in which they were recorded. In order to make easier the understanding of this work, this chapter ends with a brief overview of the

human digestive system.




2.1 WCE system details

The first prototype of the modern WCE was made in 1981. Its creator, the Israeli Gavriel Iddan, worked at the research center of the Israel Defense Forces (IDF) to design elettro-optical visors applicable on rockets. Iddan exploited the latest miniaturization technologies to create a new revolutionary system for gastrointestinal endoscopy, in which images are acquired by means of a swallowable micro-camera traveling within the digestive tract driven by the peristaltic movements. Once built an initial prototype and carried out their feasibility studies, it was performed testing on animals. In 2001, with the approval of FDA (Food and Drug Administration) and applying the CE mark , the application was extended to human people as a system to routine diagnostic endoscopy [3]. Around the capsule, Iddan founded the company “Given Imaging” [1] that aims to develop and commercialize worldwide this new technology. The system was patented under the name “Mouth 2 Anus (M2A) Given Diagnostic Imaging System.” Capsule endoscopy immediately raised great interest as it opened the opportunity to exhaustively explore the entire small intestine without any discomfort. With thousands of physicians now using the capsule as part of the initial endoscopic check-up, several articles have been written indicating this new technology suitable for some diseases as small bowel tumors, celiac disease, bleedings [4, 5].

The WCE technology is composed by means of three main subsystems: a ingestible capsule, a recording device and a workstation equipped with proprietary data processing software.

The **capsule** is disposable, which means it will not be recovered after the expulsion that naturally occurs 10 to 72 hours after the ingestion. The characteristics of the capsule may vary depending on the manufacturer. The pioneer WCE company, Given Imaging, commercializes capsules for the visualization of the mucosa in the small bowel (PillCamTM SB), esophagus (PillCamTM ESO) and colon (PillCamTM COLON). The first two kinds of

Table 2.1: The first generation of video-capsules produced by Given Imaging.

Video-capsule	Release Date	Dimensions	Frame Rate	Working Time
 PillCam™ SB	2001	11 ϕ \times 26 mm	2 fps	\sim 8 hours
 PillCam™ ESO	2004	11 ϕ \times 26 mm	18 fps	\sim 20 minutes
 PillCam™ COLON	2006	11 ϕ \times 31 mm	4 fps	\sim 4 hours

capsules have the same dimensions, even if they acquire images at two different frame rates. This is due to the different travel times of the capsule, which takes about 8 hours to go through the small intestine and 15-20 minutes inside the esophagus. All small bowel capsules have only one camera, whereas Given Imaging's esophagus and colon capsules have two of them. The idea of adding a camera comes from the need to maximize the covered surface for intestinal regions of higher diameter. It also helps to store more information in regions where the capsule travels quickly. PillCam™ COLON captures 4 frames per second and the imaging devices on either end of the capsule provide a near 360° view of the colon (Table 2.1).

The quality of WCE imaging has improved through the years. The resolution and lighting conditions are now significantly better. Given Imaging company has already reached the next generation of the capsules, which satisfies a higher frame rate maintaining the dimensions of the old generation. Another capsule distributor, Olympus [2], produces capsules based on the same size as the Given capsule but with a charged-coupled device (CCD) rather than a CMOS imager. It also provides a viewer showing real time information on the route covered by the capsule inside the patient's torso.

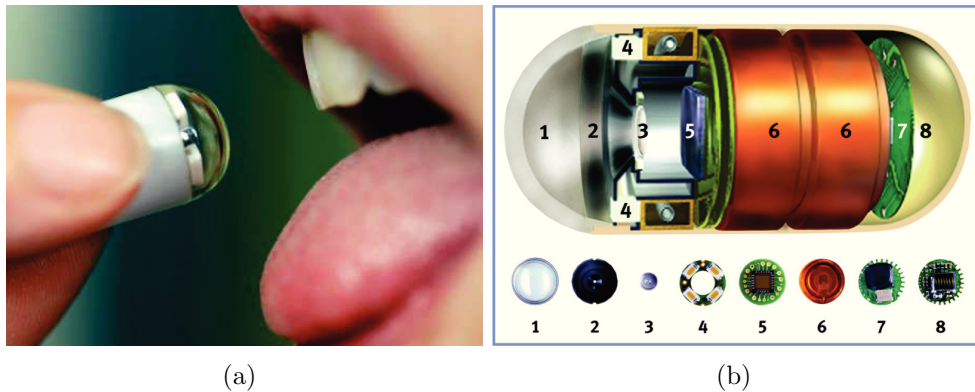


Figure 2.1: (a) Illustration of a video-capsule together with the distribution of its components in scale (b).

Figure 2.1 shows the structure of the PillCam™ SB capsule¹. It is an assembly of well-trying image acquisition components. The external case is a biocompatible plastic capsule weighting $3.7g$ and measuring $11\text{ mm} \times 26\text{ mm}$. The body of the capsule hosts the following components: an optical dome (1), a lens holder (2) with a short focal length lens (3), four illuminating LEDs (4), a CMOS (Complementary Metal Oxide Semiconductor) sensor (5), two batteries (6), an application-specific integrated circuit (ASIC) radio-frequency transmitter (7) and an external receiving micro-antenna (8). The capsule comes from the manufacturer ready to use and it starts to transmit on removal from a storage compartment, which contains a magnet that keeps the capsule inactive until use.

The **recorder device** involves the use of eight receiving antennas taped to the patient's torso, similarly to the electrodes adhesives used for the electrocardiograms. These collect the signal transmitted by the capsule and send it to the receiver carried on the patient's belt (Figure 2.2). It starts to record as soon as a signal is received from the video-capsule. The characteristics of the recorder allow the patient to wear it easily under clothing and to continue a normal daily activity. Patients are asked to avoid abrupt movements and

¹Notice that the specific Given Imaging products are used here to explain the WCE technology. After Given PillCam was introduced, many types of video-capsules have been developed and are available in the market [6, 7, 8].

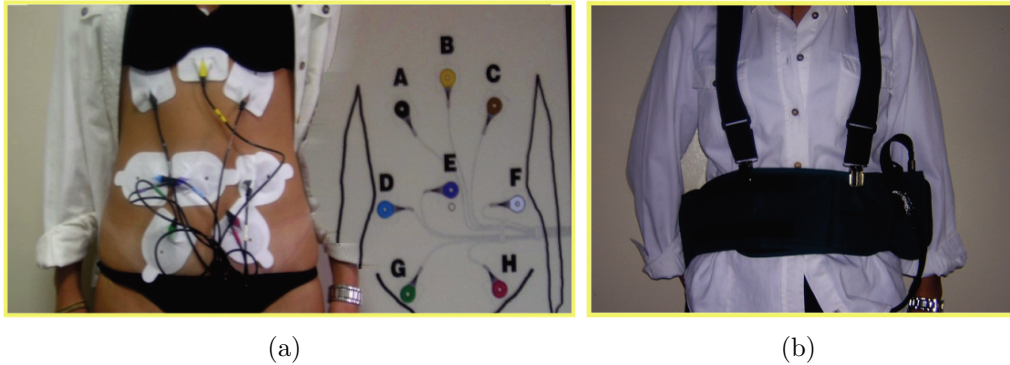


Figure 2.2: (a) The antennas array that transmits the capsule's signal to a recorder worn by the patient (b).

to constantly monitor a flashing light on the receiver for the confirmation of a good signal reception. In the last years the recorders have been improved. The capacity, battery life and reliability are now significantly better to reflect different types of capsules. They are easy to use and contain intuitive LEDs for signal reception and battery level. A typical capsule endoscopy exam takes approximatively 7-8 hours. Once the exam is finished, the patient comes back to ambulatory to deliver the recorder containing all the images captured by the capsule and wirelessly transmitted.

The workstation is a dedicated computer equipped with a proprietary **data analysis software**. It allows to watch the entire examination, with a special utility enabling to get quickly any image within the video. The physician reads the video in one of several formats, captures and labels the salient information, and then prepares a final report. As already mentioned, this is one of the main drawbacks of endoscopy through video-capsule. Notice that a good diagnosis requires up to two hours, and it highly depends on the physician's experience. This process is also so exhausting that the physician rarely performs two consecutive diagnoses. At the end of the analysis all the annotated information are automatically saved. These findings files can be saved on a CD, a DVD or any other storage device and then sent to colleagues for consultations. For further details on using this software, refer to Section 2.4.

2.2 Benefits and risks of capsule endoscopy

Since its introduction, more than 1.000.000 patients worldwide have benefited from the capsule endoscopy. As with all new technologies, the practical use of the capsule led to the introduction of improvements to the diagnostic system that, in turn, has opened new fields of clinical use. Already in the early studies it was possible to assess the operational characteristics and the limits of the capsule:

- **Usability:** It is much less invasive than traditional endoscopy, since the patient simply has to swallow a vitamin-size capsule, which will be expelled in the normal cycle through the defecations. The application of the system (antennas, recorder, batteries) to the patient is very easy and can be carried out, with a minimum of training, even by unskilled staff.
- **Tolerability:** The majority of the patients succeeds in swallowing the capsule with some sips of water. The application of the antennas on patient's torso has not resulted in complaints by patients.
- **Completeness of the examination:** As already mentioned, the capsule is strongly indicated to the exploration of the small intestine. It sometimes allows to capture images of the esophagus, but the rapidity of oesophageal transit rarely allows to capture significant findings. In such cases it is suggested to use the Given PillCam ESO: it contains an imaging device at both ends of the capsule and take up to 18 frames per second as it passes through the esophagus. Nowadays, however, capsule endoscopy cannot exhaustively replace the use of standard endoscopic procedure; indeed, it is often complementarily used with other examinations. Since the capsule has not therapeutic capabilities, any abnormalities detected by the capsule must be further investigated by the standard endoscope. This has the proper tools for the extraction of intestinal tissue destined to a later histological investigation. Furthermore, the movement of the capsule within the digestive tract is passive and driven by peristalsis. The recorded area is hence unpredictable, it

is patient dependant, and it is not conceivable the employment of the capsule in place of the standard endoscopy.

One of the main concern, although it rarely appears, is related to the capsule retention when it is not naturally excreted in the feces within two weeks after the ingestion. If it occurs, it is necessary to remove it through the surgery. Capsule endoscopy is hence contraindicated in patients with known or suspected intestinal stenosis or with the presence of severe deformities of the digestive tract. Recently, a “patency” capsule that does not require any preparation has been introduced in the market. The utility is to scan the bowels to verify an adequate patency of the gastrointestinal tract in patients with known or suspected strictures prior to administration of the video capsule in safety and tranquillity. This capsule is made of specific materials decomposing with the contact to the intestinal contents in a few days after the ingestion. It also contains a Radio Frequency Identification (RFID) tag to determine capsule location.

Other contraindications to the application of the video-capsule refer to patients which have suffered previous invasive surgery on the abdomen and when there is the presence of pacemakers or other electrical medical equipment. This is due to the possible interferences between those systems and the WCE radio transmitter. Some technical malfunctions related to the capsule have been reported but were rarely significant [9]. The increasing interest in this technique and the technology improvement would make these issues occurring less frequent.

The most critical limitation concerns the discharge time of the images from the portable recorder and the long annotation time that each exam needs from a trained specialist. He/She may spend up to one or more hours to gather the relevant information for a proper diagnosis. This greatly limits the use of the capsule as a diagnostic routine tool. Such shortcoming may be overcome if the WCE video is automatically segmented into shorter videos, each one relative to a different trait of the bowels, and if reliable automatic annotation tools are available to the clinicians. Unfortunately, the goal of automatically producing a summary of the whole WCE video remains yet unaccomplished. Tools to extract semantic information from such videos,

such as the one presented in this work, are relevant research products for applied Pattern Recognition investigators.

2.3 Typical capsule endoscopy images

Wireless Capsule Endoscopy produces images of the digestive tract, covering a circular 140° field of view. An average exam has around 50.000 images where 1000 are captured in the gastrointestinal tract entrance, 4000 in the stomach, 30000 in the small bowel and 3000 in the large intestine. The captured frames have three 8 bit color planes with a 256×256 pixels resolution, rendering a circular area of 240 pixels of diameter. The black area surrounding each rendered frame usually contains some further information, like the exam's date and the patient's name (Figure 2.3).

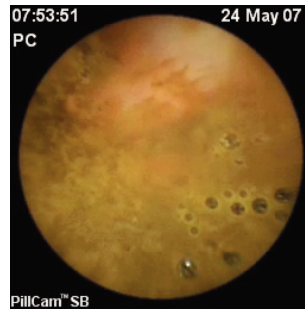


Figure 2.3: A typical Wireless Capsule Endoscopy frame.

Each incoming frame can be visually classified according to the characteristics of the intestinal mucosa and other typical elements that may be present into the bowels, like bubbles, bleedings, residuals, ulcers, etc. The physician recognizes these events mainly using color and texture pattern, determining the status of the intestinal mucosa. Figure 2.4 illustrates sample images of healthy regions and organic lesions of the gastrointestinal tract. In particular, Figure 2.4(a) shows a detailed view of the normal mucosa of the small bowel. It is an intestinal region with uniform pink hue. Bleeding is defined as the flow of blood from a ruptured vein into the digestive tract. The visual feature used to characterize this scenario may rely on the intensity of the red

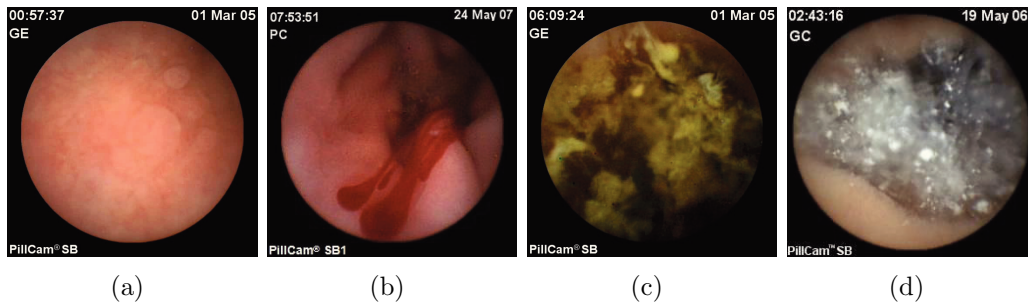


Figure 2.4: Different examples of capsule endoscopy scenarios. (a) Normal small bowel mucosa. (b) Bleeding. (c) Residual. (d) Intestinal juices.

color component (Figure 2.4(b)). Instead, green color is usually related to the presence of fecal materials (Figure 2.4(c)). In WCE, the good visibility of the internal tissue is sometimes obstructed by the intestinal juices, which can be visualized as a turbid liquid accompanied by bubbles or other elements related to the flow of different gastric juices (Figure 2.4(d)).

It should be noticed that the gut is not motionless; the physiological motion peristalsis may reverse or incline the capsule recording a variety of orientations of the scene. In addition, it is needed to consider external factors such as the lighting of the capsule which can sometimes falsify the perceived colors. The number of scenarios in which a certain event can be recognized is hence very impressive.

2.4 Manual annotation

Once the examination is finished, the patient delivers the data recorder containing the images captured by the capsule. A workstation with proprietary software is used by the physician for the analysis of the video. This means that the physician needs to view the full 50 thousand images², annotate all the relevant ones, and create a final medical report with the summary of the conducted investigation. The expert also includes all the diagnostic conclusions and any checks to be performed for the monitoring of certain diseases.

²The number of collected frames may range depending on which type of examination is performed and the frame rate used by the capsule.



Figure 2.5: Rapid Reader exam annotation software developed by Given Imaging.

Performing correctly the analysis is difficult, requires trained staff and time to perform the needed analysis.

Figure 2.5 shows a snapshot of the visualization tool provided by Given Imaging: Rapid Reader. Two main motivations lead us to adopt this software. Considering that we have a dataset of images coming from Given capsules, the only way to handle this data is with the dedicated software. It can also be downloaded for free directly from the manufacturer's site.

The RAPID (**R**eporting **A**nd **P**rocessing of **I**mages and **D**ata) software suite enables efficient management of capsule endoscopy studies from initiation, through review and analysis, to report generation. With solutions for every capsule endoscopy workflow, this software suite provides multiple reading modes, advanced analysis features that aid in image interpretation, intuitive report generation, convenient study management, and network connectivity. The physician can review the WCE video by using all the available features

and utility provided with the application software. Video images can be viewed in single or mosaic format and with different frame rates. This depends on the experience of the physician; specialized users tend to display multiple images at once while maintaining a higher reading frame rate.

The software also contains three important utilities:

- The first is a **time bar** that allows the doctor to understand the context of a specific intestinal image. This bar contains the average color of the images to which it refers. In this way it is possible to track the movement of the capsule and its travel times through each intestinal organ.
- The doctor is facilitated in the preparation of the report by the presence of a comprehensive **atlas**. This provides side by side comparison of an image in a case currently under review with atlas reference images.
- Rapid Reader software includes a **Suspected Blood Indicator (SBI)** designed to detect bleedings in the video. However, this tool has been reported to have insufficient specificity and sensitivity. This means that it may display a high number of false positives but it is useful for capture regions with active bleedings.

2.5 The digestive tract

In this section we give some useful information to the reader about the human digestive system. This is done to better understand the classification tasks addressed in this work and how these differ according to different organs of the intestinal tract. A simplified description of the human gastrointestinal tract appears in Figure 2.6.

The digestive system, also known as the gastrointestinal (GI) system, can be seen as a long tube (about 4-7 meters) that passes through the body, starting at the mouth and ending at the anus. It is capable of absorbing the nutritional contents from the ingested food eliminating the waste out of the body. The function and the visual appearance of each different section of the

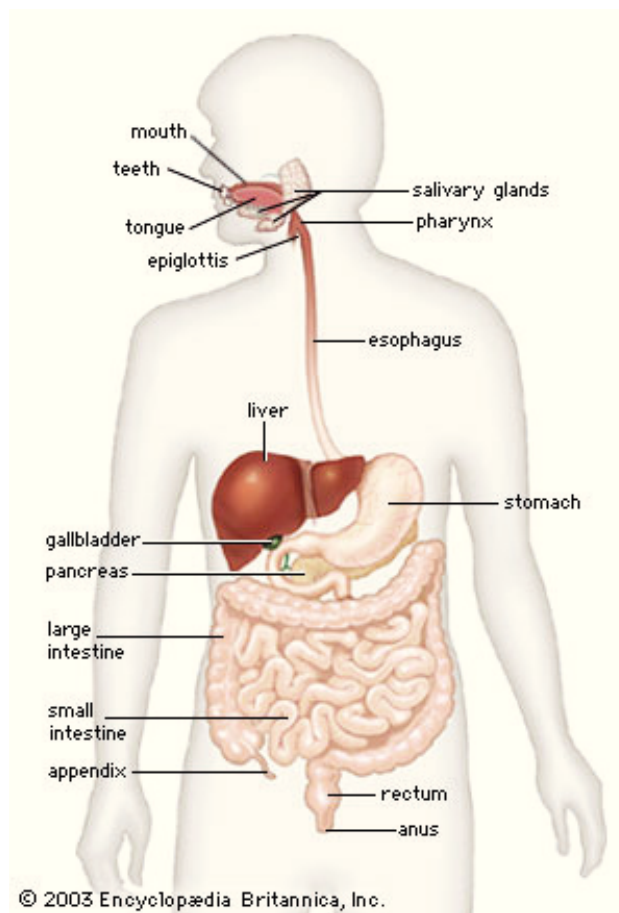


Figure 2.6: A schematic illustration of the human GI tract.

gut highly depends on the physiological task to which is part is devoted. A first distinction is generally done between the upper GI tract and the lower GI tract. The first one is composed by the oral cavity (mouth and pharynx), the esophagus, and the stomach. Typically, these portions of the digestive system are viewed using standard probe-based endoscopic procedures. The stomach is a big bag covered with a thick mucosa membrane containing gastric juices. In its relaxed state contains several longitudinal folds and in the pylorus, the terminal region of the stomach, the diameter is about two centimeters. Automatically locating the pylorus is of great advantage because it provides the expert with the point at which food passes into the duodenum, the first part of the small intestine.

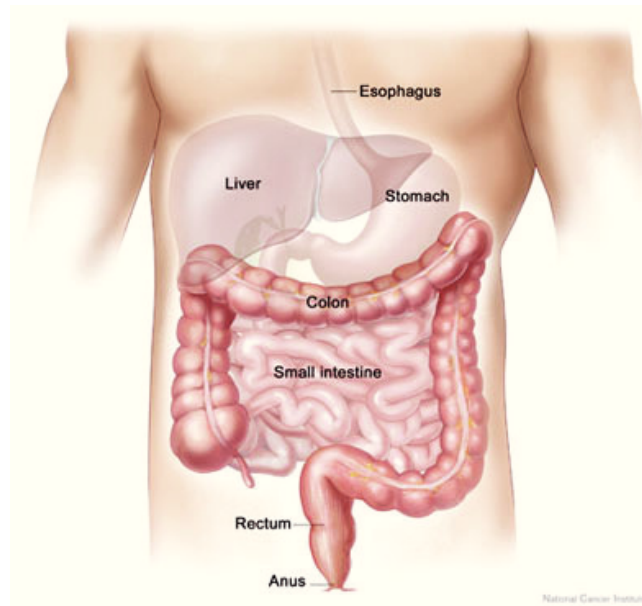


Figure 2.7: The lower GI tract.

The lower GI section is hardly practicable as it is much longer and articulated; it comprises the small intestine, large intestine, and anus (Figure 2.7).

The small intestine presents three different areas: duodenum, jejunum and ileum. The main duty of the duodenum is to continue the digestion done by the stomach. The next part of the small intestine is called the jejunum, and the third is called the ileum. Except by close internal histological inspection, these two parts cannot be readily separated as they present a similar visual appearance: the intestinal walls are plain in the relaxation state, but they contract creating folds during the motility activity. The ileum has a paler color, and tends to be of a smaller caliber as well. Together the jejunum and the ileum contribute more than 15 feet to the small intestinal length. The ileum ends by opening into the large intestine, or colon, via the ileocecal junction.

The last region of the intestinal tract is the large intestine which is about 150 *cm* long and 6 *cm* in diameter. It does not contain many folds as the small intestine and the larger diameter makes the diagnosis difficult for capsules designed for the small intestine.

Chapter 3

Literature review

The previous section has provided a description of WCE as a technological advancement in the area of diagnostic endoscopy. The increasingly clinical relevance is evidenced by different studies that compare the examination through the video-capsule with the traditional endoscopic procedures. It is quite clear that capsule endoscopy performs better than push enteroscopy in diagnosing patients with difficult gastrointestinal bleedings [5],[10]. It is also commonly used in other clinical conditions, such as the detection of Chron's disease in the small bowel [11],[12], celiac disease [4], small bowel polyposis and tumors [13]. Sometimes is used to study the impact of drugs on the gastrointestinal tract [14]. Moreover, children can benefit from this technology as well as adults [15].

The main issue is related to the final report of the examination. A considerable amount of time is required to view and interpret the many thousand of images produced during the examination. This is a difficult and tedious task that requires a qualified staff.

Since the vision is the main feature of an optical system such as the video capsule endoscopy, Computer Vision techniques may help to automatically select and detect the salient information enclosed in WCE data. Although the research conducted on this new endoscopic technology is still in an introductory phase, a significant number of papers have already been published. The application of Computer Vision in capsule image analysis can be di-

vided in two categories. The first considers the topographic segmentation of a WCE video into meaningful parts such as mouth, esophagus, stomach, small intestine, and colon. Regarding the second category, there are several works which seek to identify clinically relevant video events. Some instances include the automatic detection of bleedings, intestinal juice, intestinal contractions, ulcers.

In this chapter we offer a brief survey of research related to the classification of images extracted from WCE videos.

3.1 Topographic segmentation

One of the main issue in wireless capsule imaging is the creation of a map of the data recorded by the capsule during the navigation through the gut. To this aim, topographic segmentation performs a segmentation of an endoscopic video into shorter videos, each one relative to a different trait of the gut. Some intestinal diseases may reside in a specific segment of the intestine; each digestive organ thus requires a different level of attention by the clinical viewer. A comprehensive map of the examination enables a medical expert to browse to a particular areas of interest, making the analysis of the examination faster.

Most of the work found in literature tend to split the endoscopic video into four main sections (for a better understanding see Section 2.5):

- **Entrance:** Once it is activated, the capsule is outside the body for no more than few seconds. Then, it is swallowed by the patient and it quickly reaches the esophagus until the esogastric junction separating this from the stomach. This subset of data is clinically irrelevant because the capsule travels very fast in these areas that can adequately be observed using traditional probe-based endoscopic procedures. From an Image Processing perspective, it is possible to find images with several color and texture variations: we can see the capsule cover, outside world, teeth, tongue, etc.

- **Stomach:** This area begins in the esogastric junction and ends in the pylorus. Although it is clinically relevant, it is difficult to find relevant events since the peristalsis may reverse or incline the capsule due to the higher diameter of this tract. The images in this area are usually light red and smooth. Finding the pylorus in the video can be difficult and time-consuming, even for an experienced viewer, as visually the stomach tissue in the pyloric region and the tissue at the beginning of the intestine appear very similar.
- **Small Intestine:** This is the region in which the capsule has the most significant clinical impact. The small intestine is the longest region of the gastrointestinal tract. This tubular section usually contains semi-digested foods, intestinal juices, enzymes. It is divided from the colon by the ileocecal valve. Annotating this boundary is even more difficult because intestine and colon tissue are very similar and are often contaminated with faecal residuals that occludes the camera view.
- **Large Intestine:** The last topographic section encountered by a WCE begins in the ileocecal valve and normally ends when the capsule's battery runs out. This area suffers from very low visibility due to the high concentration of food and faecal material. It does not contain as many folds as the small intestine, but the larger diameter leads the capsule to freely move making the diagnosis difficult for capsules designed for the small intestine.

The topographic segmentation task is roughly equivalent to the search of those boundaries in the video. Since 2001, a considerable number of works have been published regarding this task.

The authors in [16] propose a technique to perform the boundary detection task based on color change pattern analysis. When a capsule travels around a boundary between two different digestive organs, the corresponding color signal has a sudden change. This methodology characterizes the contractions of WCE videos using energy function in a frequency domain. They segment a WCE video into events by using a high frequency content function. The detected boundaries event indicate either entrance in the next organ or unusual

events in the same organ, such as bleedings, intestinal juices, and unusual capsule movements. It is hence possible that boundary events may contain other smaller events representing something else. The authors classify these events through a threshold-based correlation rule into higher level events that represent digestive organs. The experimental results indicate that a high percentage (76%) of detecting correct boundaries events and a precision of 51% have been achieved. The methodology manages to detect the most of stomach and duodenum, but the accuracy in the ileum and cecum is worse. A relevant series of papers performs an automatic gastrointestinal tissue discrimination resulting in the segmentation of various intestinal organs [17, 18, 19, 20, 21, 22, 23]. In these papers, feature extraction procedure is performed in the same way: the authors create a feature vector using color and texture information. To this aim, images are initially converted in *HSI* color space. They derive a color features from Hue Saturation chromaticity histograms, compressed using a hybrid transform, incorporating the Discrete Cosine Transform (DCT) and Principal Component Analysis (PCA). Because of the abrupt intensity changes in WCE images, the intensity component is removed to achieve intensity invariance and data size reduction. In [17] a second feature combining color and texture information is derived using Local Binary Pattern (LBP). Having extracted feature vectors, the next stage involves classifying them as belonging to a specific digestive tract. There are several classifiers which can perform this task. In [19] the system is trained to detect mouth/esophagus and stomach, stomach/intestine, and intestine/colon using k-nearest neighbor (KNN) and Support Vector Machine (SVM) classifiers. The work in [20] is similar to the previous one; additional regions have been discriminated both in stomach and intestine. Histograms built using the entire image may contain visual contamination present in the image. Some examples of such noise is the presence of bile, saliva, food remains, air bubbles and so forth. In order to minimize the affect of noise in the image, the authors divide the WCE image into 28 sub-regions and process only those regions where tissue is clearly visible. They derive five parameters for each of the sub-images (Mean Intensity, Saturation, Hue, Standard Deviation of Intensity and Hue). Then, each sub-image is tested and it is

discarded if exceeds the range of reference values for visually clear images of stomach or intestine tissue.

Coimbra et al. [24, 25, 26, 27] deal with the task of topographic segmentation by using a novel visual descriptor called MPEG-7 [28]. This defines a variety of visual descriptors for multimedia content, including audio, speech, graphics and their combination. In [24] the authors use this descriptor adapted to the WCE specific scenario. The final segmentation is based either on Bayesian or SVM classifiers. In particular they trained four SVMs classifiers, one for each boundary (esogastric junction, pylorus, ileo-cecal valve), determining thereby the belonging topographic section of each frame. Despite good results have been achieved in [24, 25], the authors suggest to use content features with context information [29]. Such new information may include the approximated capsule spatial location inside the body with the relative capsule velocity.

All these works disregard the computational cost to perform the segmentation. Notice that some capsules manufacturers [2] offer original viewer to physician providing real-time examination imagery. It is hence highly desirable that these computer-assisted tools can run on simple portable hardware that might be incorporated, for example, next to the portable hard-drive carried in the patient's belt during the procedure. To this aim, in a more recent paper Coimbra et al. [30] show how a compressed domain color information can be used to perform topographic segmentation as well as algorithms using fully decoded images saving about 20% of the computational cost.

3.2 Event detection

With the video segmentation into coherent intestinal sections, the expert can easily access to the images taken in a specific intestinal tract. The next step consists in automatically detect some typical intestinal scenarios in order to reduce the analysis time by the expert to do a diagnosis. Only a small amount of interesting images for diagnostic purposes are indeed usually spread in thousands of useless images. Notice that a higher amount of false positives than of false negatives is typically preferred for this kind of applications. The

presence of a high number of false positives result in more time spent by the expert to do a diagnosis. Losing a rightful event is a worse event because it means to miss relevant information with the resulting inaccuracy in the final report.

3.2.1 Intestinal contractions

Dysfunctions of the intestinal motility are often related to certain disorders that may occur with different symptoms [31]. The analysis of intestinal contractions in the small intestine, in terms of number, frequency and distribution, represents one of the methods with greater clinical significance. An intestinal contraction involves a sequence of frames in which the intestinal lumen shrinks, tending to the maximum closure, and then expands again. The typical appearance hence consists in a dark area surrounded by the typical rays that muscular tone produces due to the folding of the intestinal wall. In [32] the authors propose a technique based on anisotropic image filtering and efficient statistical classification of contraction features. The procedure to detect the typical star-shape of a contraction is accomplished in three steps. The skeleton of the wrinkle pattern is extracted. Then, it is verified whether the point at which all the rays converge corresponds to the closure of the intestinal lumen. Finally, a set of descriptors were estimated taking into account the radial organization of the wrinkle skeleton around the intestinal lumen. Classification is performed by using a SVM classifier with radial basis function. The system reaches a sensitivity of the order of 90.84% and a specificity of the order of 94.43% respectively.

3.2.2 Intestinal juices

In many images the information needed for a correct diagnosis is sometimes obscured by intestinal elements, such as gastric juice, bubbles, residuals. These intestinal elements are constituted by turbid liquids with color ranging from green to white. They often are visualized together with bubbles or other artifacts related to the flux of different fluids into the gut (Figure 3.1).

The authors in [33] point out the the most relevant feature of the intestinal

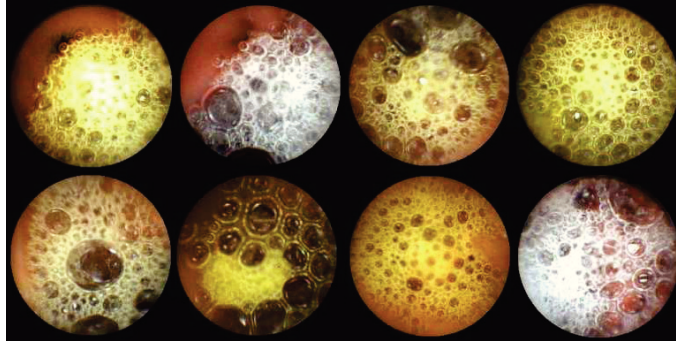


Figure 3.1: Examples of endoscopic frames showing intestinal juices.

fluids is the presence of bubbles of different sizes and circular shapes. To characterize these items they rely on the use of a Gabor filters bank of 16 units with orientations 0° , 45° , 90° , 135° and standard deviations $\sigma = 1, 2, 4, 8$. By using a threshold-based mechanism, they obtain a binary image in which the intestinal fluids are emphasized. Those images with detected region of bubbles greater than 50% of the useful visualization area are excluded. It can be observed that an overall reduction in visualization time about 23% is achieved.

3.2.3 Bleeding detection

Bleeding in the digestive tract is often a symptom of some diseases, rather than a disease itself. The cause of bleeding may not be serious, but locating the source of bleeding is critical. The Given proprietary software provides an automatic image analysis tool called Suspected Blood Indicator (SBI), which is devoted to find in the video areas with active bleedings. However, this tool has been reported to have insufficient sensitivity and specificity [34]; to be safe, the physician must continue to manually check for bleedings.

To overcome this problem, in [35] a technique to automatically detect the bleedings regions using the expectation maximization (EM) clustering algorithm and Bayesian information criterion (BIC) is proposed.

When it is performed a visual inspection of gastrointestinal images, one of the most salient feature indicating the presence of bleeding is a deeply red

appearance or the presence of dark red regions. Finding a red dominant color in the gut is very common, but the red in non-bleeding regions usually appear with lower color saturation. The authors in [36] exploit this idea and propose a two-steps bleedings detection algorithm. The first step provides an efficient block-based discrimination of the input frames that contain bleeding features from those that do not correspond to bleeding. The second step refines the initial classification and increase its reliability using a pixel-based saturation-luminance analysis. To make the classification more complete, images are categorized into several levels of bleeding activity or as non-bleeding. A 4-level bleeding classification is applied: level 0 indicates non-bleeding and level 3 indicates highly intensity bleeding. The system sensitivity achieved is 88.3%.

Recently Li et al. [37] propose a method to detect either bleeding and ulcer by means of chromaticity moments, which make use of the Tchebichef polynomials and the illumination invariance provided by the *HSI* color space. To reduce effects of visual contaminations such as bubble, fecal material and the dark regions that occur often in this kind of images, the authors divide the endoscopic frame into a grid of 36 non-overlapping blocks calculating six chromaticity moments for each one. The blocks are finally classified using an MLP (multilayer perceptron) Neural Network with 4-fold cross-validation.

3.2.4 Anomaly detection

In this section we provide a brief overview of published works related to the detection of abnormal lesions in endoscopic images. In this way, one of the first Image Processing studies was conducted by Boulougoura et al. [38]. The authors propose to use a feature vector composed by nine measures (standard deviation, variance, skew, kurtosis, entropy, energy, inverse different moment, contrast and covariance) extracted from histograms of six channels (R,G,B,H,S,V). The implementation of an advanced neural network scheme and the concept of fusion of multiple classifiers dedicated to specific feature parameters have been adopted to classify the images as normal or abnormal. The detection accuracy achieved by this system is 100%. However, at the

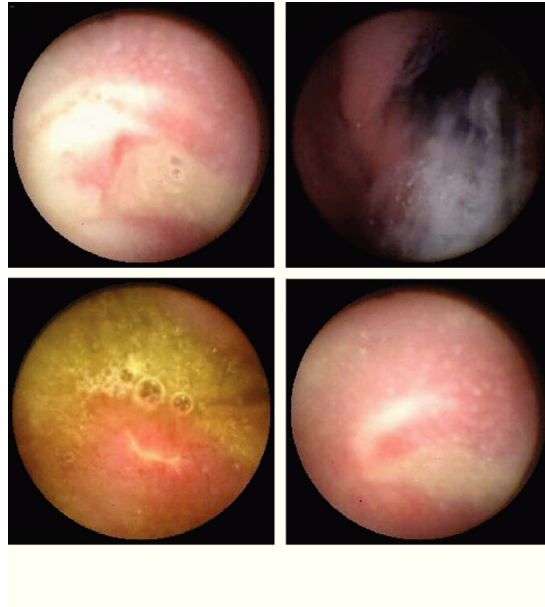


Figure 3.2: Examples of endoscopic frames showing ulcers.

time of this paper the system was evaluated using only 73 capsule images, which were split into the training set and the test set, and therefore it is difficult to draw conclusions about its potential application in a diagnostic station.

Li et al. [39] propose a new feature extraction scheme based on the use of curvelet transformation and LBP to discriminate ulcer regions from normal regions (Figure 3.2). The traditional wavelets transform extracts directional details capturing only horizontal, vertical and diagonal activities in an image, and these three directions cannot in general provide enough directional information in images. To this aim, curvelet transform is employed as a new multi-resolution analysis tool. The basic idea is to represent a curve as a superposition of functions of various lengths and widths obeying a specific scaling law. Taking into reference 2D images, this may be obtained by decomposing an image into wavelet sub-bands. Each sub-image of a given scale is then analyzed with a ridgelet transformation, another type of tool for multi-resolution analysis. It should be known that capsule endoscopy images suffer from sudden changes in lighting due to the movement of the capsule and the

limited range of illumination inside the digestive tube. The authors propose to use LBP after they applied curvelet transformation to images. In this way they obtain robust performance to illumination variations. By using uniform LBP histogram, they obtain six statistical measurements of the histogram as features of texture in order to reduce the number of features. These features are standard deviation, skew, kurtosis, entropy, energy and mean of the histogram. To reduce effects of visual contaminations such as bubble, faecal material and dark regions that occur often in capsule endoscopy images, they divide each frame into small patches extracting textural features from each one of them. To verify the performance of this features extraction scheme, the authors deploy MLP neural network and SVM to demonstrate their power in differentiating normal regions and ulcer regions in capsule endoscopy images. The best results are obtained using the MLP classifier and the YCbCr color space. With these parameters the system achieves an accuracy of 92.37%, a specificity of 91.46% and a sensitivity of 93.28%.

More recently, Bejakovic et al [40] present a method that uses color, texture and edge features to detect lesions (in particular Chron's disease) in capsule endoscopy images. They use MPEG-7 visual descriptors and Haralick texture features. This includes MATLAB adaptation of dominant color (DCD), homogeneous texture (HTD) and edge histogram (EHD). Haralick features include angular moments, contrast, correlation, and entropy measures, which are computed from 1 pixel co-occurrence matrix. They used SVM to classify images into three categories: lesion, normal tissue, and extraneous matter (food, bile, stool, air bubbles, etc). The dataset used to evaluate their system is composed by images selected from ten studies. For each study only 10% of the data was used to train the classifier; the remaining data was used for validation. Over the ten studies, lesions could be detected with an accuracy rate of 96.5%, normal tissues 87.5% and extraneous matter 87.3% using dominant color information alone.

Finally, Li et al. [41] develop a computer aided system to diagnose small bowel tumors. They propose a new textural feature built by using wavelet and LBP. Notice that tumors exhibit great variations in color, size and shape, so a single classifier may not be discriminative enough to make a right decision

about status of these difficult images. To this aim, the system is evaluated using an integration of KNN, MLP neural network and Support Vector Machine. The database used for the classification consists of 600 representative small bowel tumor images and 600 normal images previously labeled by the physician. Comparative experimental results show that this scheme achieves a promising performance for small bowel tumor detection.

Chapter 4

Information Theoretic Method

Information Theory aims to identify the theoretical concepts for the study of problems related to the transmission, reception, processing and storing information. Although at first sight it seems to be a very specific application, this theory has important implications (and applications) in many areas. Information Theory is usually dates back to 1948, when Claude Shannon published “A mathematical Theory of Communication” in which he introduced for the first time a systematic study about information and communication. He also formulated the key concepts of the theory, such as entropy and mutual information and introduced the fundamental laws of data compression and transmission. Entropy is a measure that quantizes the information contained in a random process. Mutual information is a measure of the information contained in one process about another process. However, Shannon’s entropy is relative to some probability distribution generating data. In many cases such a distribution is unknown or does not even exists. To this aim, Kolmogorov complexity has been connected with Information Theory and proved to be closely related to Shannon’s entropy rate of an information source. In both theories, the amount of information in an object may be defined as a function of the length of the object’s description. In the Shannon approach, however, the method of encoding objects is based on the presupposition that the objects to be encoded are outcomes of a known random source; it is only the characteristics of that random source that determine the encoding, not

the characteristics of the objects that are its outcomes. In the Kolmogorov complexity approach we consider the individual objects themselves and the encoding of an object is a short computer program (compressed version of the object) that generates it and then halts [42]. Kolmogorov complexity is a measure that answers the question: how random is an individual bit string or message? Complexity is in this case intended to define the amount of information contained in a particular message in terms of the number of bits necessary to describe it. More random objects would require more bits to describe them, so the Kolmogorov complexity is also a measure of how random a particular message is. In contrast, the Shannon entropy is an answer to the question: how random is an entire distribution of messages overall? Entropy measures the expected amount of information contained in any given message within that distribution.

In this chapter, some basic notions of Information Theory and Algorithmic Information Theory describing an absolute information-theoretic distance between bit strings are presented.

4.1 Entropy

Given a generic event E , how can we define the informative content of this event? Shannon starts from the consideration that the observer should have some ideas of the probability that such an event occurs. The basic idea is that the information has to deal with uncertainty: more the observer is surprised to see a symbol, the more the level of informative content will be high (and viceversa). Suppose we have a set of possible events with *a-priori* distribution probability of occurrence $P = (p_1, p_2, \dots, p_N)$. To numerically estimate how much information is gained on average or the degree of uncertainty with a given function $H(P)$, the following basic properties must be guaranteed:

- The function must exist, i.e., it must be possible to associate a link between the numerical uncertainty of a probability distribution and the real numbers.

- $H(p_1, \dots, p_N)$ is continuous in p_1, \dots, p_N . This means that to small variations of P correspond small variations of H .
- If all the p_i are equal, $p_i = \frac{1}{N}$, then H should be a monotonic increasing function of N . In other words, as the number of events increases, if they are equally probable, the associated uncertainty also increases.
- H must guarantee the additive property: if a event is splitted into two successive events, the original H should be the weighted sum of the individual values of H . A clarifying example of this property is shown in the figure 4.1.

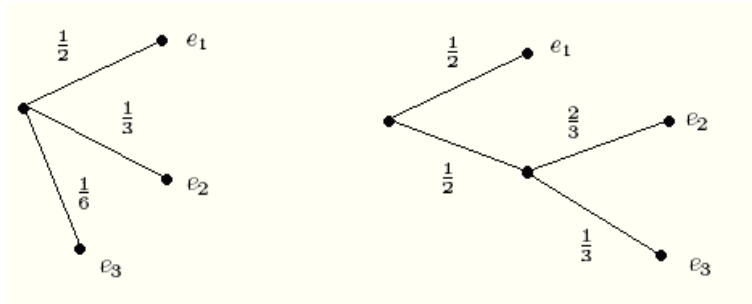


Figure 4.1: Grouping property of the entropy.

We can think of $P = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$ as being generated in two successive choices $p' = (\frac{1}{2}, \frac{1}{2})$ and $p'' = (\frac{2}{3}, \frac{1}{3})$. Thus, the entropy of P must be equal to entropy of the first step in the decomposition process, plus the weighted sum of the entropies of the second step:

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}H\left(\frac{2}{3}, \frac{1}{3}\right) \quad (4.1)$$

The coefficient $\frac{1}{2}$ in the entropy of the second step means that the second choice only occurs half the time.

The only H satisfying the four above assumptions is of the form

$$H = -K \sum_{i=1}^N p_i \log p_i \quad (4.2)$$

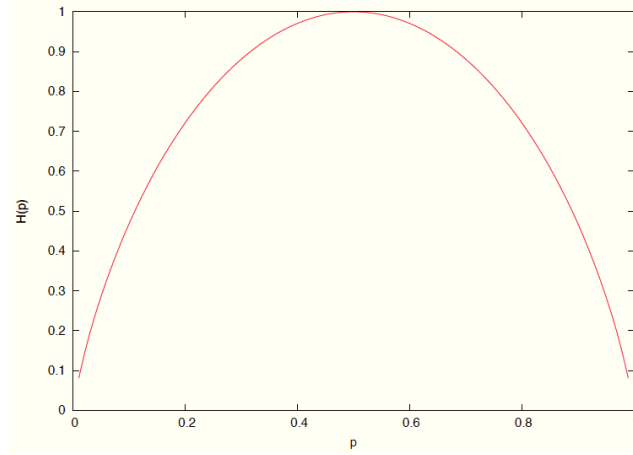


Figure 4.2: Entropy of a binary source.

where K is a positive constant. We interpret $0 \log 0$ as equal to 0, which follows logically from $\lim_{x \rightarrow 0} x \log x = 0$. When $K = 1$ and the logarithm is \log_2 , information is measured in *bits*. If the probability distribution has only two entries with probability p and $1 - p$ respectively, entropy will be equal to

$$H_{bin} = -p \log p - (1 - p) \log (1 - p) \quad (4.3)$$

The graph of the function H_{bin} is shown in Figure 4.2. It is a concave function with null values for $p = 0$ and $p = 1$; this occurs when the variable is not random and there is not uncertainty. Similarly, with the value $p = \frac{1}{2}$ uncertainty is maximum.

Some other relevant properties of the entropy are:

- H is symmetrical with respect to the probability vector P from which it depends, in the sense that if it is performed a permutation of the elements on the vector P , entropy does not change.
- $H(P) = 0$ if and only if all the probabilities except one are zero.
- $H(P) = \log n$ when all the probabilities are equal. This is the case in which the informative content is maximum.

It is worth mentioning that, in Information Theory, the concept of entropy is closely linked to that of compression. If the content of a message (for

example a simple string) is uncorrelated with each other, there is no way to compress the message without losing information and this is the case where entropy is maximum. Conversely, if some parts of a message are logically related to other, the resulting entropy will be smaller and consequently the message can be compressed without loss of information. The entropy of a set of data is hence directly related to the amount of information that it contains and provides a theoretical bound on the amount of compression that can be achieved [43].

4.2 Kolmogorov complexity

One of the most difficult and representative problem in Computer Science is represented by the problem of Complexity, i.e., the measure of the computational resources required to perform a computation. There are two different categories of complexity:

- Static complexity, related with the structure of the program and its size.
- Dynamic or computational complexity, typically divided in time complexity and space complexity.

In the context of static complexity, the Kolmogorov complexity, independently introduced by R.J. Solomonoff in 1964 [44], by A.N. Kolmogorov in 1965 [45] and by G.J. Chaitin in 1966 [46] is very relevant to the development of an Information Theory based on the length of the codes.

Kolmogorov complexity (or algorithmic entropy) of a string x , defined as $K(x)$, is the length $l(p)$ of the shortest program p that runs on a universal computing device (a Universal Turing Machine φ) and produces the string x as output. Intuitively, it represents the minimum amount of information required to generate the string x from some effective process. For example, the string consisting of n equals symbols is extremely simple, since the shortest program that generates it has only to print n times the same symbol.

Mathematically, the Kolmogorov Complexity is defined as follows [47]:

$$K(x) = \min_{\{p|\varphi(p)=x\}} l(p) \quad (4.4)$$

Intuitively, the above equation describes a competitive selection of the shortest program, denoted p^* , from an unbounded set of competing programs $\{p_0, p_1, \dots\}$, each one capable of producing the desired output x . The unlimited nature of this competition ensures that the winning model can be more efficient to find all the structural regularities in the specific string. Unfortunately, this also implies the lack of a guarantee that this competition always produces a result. Despite the great interest shown in the scientific community, actual results about the Kolmogorov complexity have not been achieved. The difficulty is related to the impossibility to evaluate the function K for a given string x , because it is not Turing computable. The uncomputability of Kolmogorov complexity has motivated several authors to seek useful approximations. A very good reference is the classic “Vitany trilogy” [48, 49, 50] in which practical approaches for approximating the Kolmogorov complexity and the related notion of Algorithmic Information Distance using compression algorithms are presented.

4.3 Algorithmic Information Distance

While the Kolmogorov complexity is a measure of the information contained in a individual object, a similar concept for the information distance between two individual objects is required. In this section we consider the problem of the definition of a distance D between two generic string x and y . This distance measure must satisfy the following requirements:

- Positivity: $D(x, y) \geq 0$ ($D(x, y) = 0 \leftrightarrow x = y$)
- Symmetry: $D(x, y) = D(y, x)$
- Triangle inequality: $D(x, z) \leq D(x, y) + D(y, z)$

Intuitively, it is possible to calculate the similarity between two strings x and y as the length of the shortest program that computes x from y and viceversa. Bennett et al. [48] defines this measure, denoted $E(x, y)$ as:

$$E(x, y) = \max\{K(x|y), K(y|x)\} \quad (4.5)$$

where $K(x|y)$ is the conditional Kolmogorov complexity of a string x related to string y defined as the length of the shortest program to compute x if string y is provided to the universal computer as an auxiliary input. Sometimes the distance analysis requires a normalized metric. This requirement can be satisfied by the universal similarity metric defined by Li et al. [50], known as *NID* (Normalized Information Distance) and mathematically represented by the following formula:

$$NID(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \quad (4.6)$$

NID is an universal distance measure for objects of all kinds. It is also based on Kolmogorov complexity and thus uncomputable, but in the Vitany trilogy [48, 49, 50], the authors propose a way to approximate it. Compression algorithms can be used to approximate the Kolmogorov complexity if the objects have a string representation. Let $C(x)$ the size of the compressed version of the string x and $C(x, y)$ the size of the compressed version of the concatenation of x and y . We can rewrite the *NID* to obtain the Normalized Compression Distance (NCD) :

$$NCD(x, y) = \frac{C(x, y) - \min(C(x), C(y))}{\max(C(x), C(y))} \quad (4.7)$$

Standard Compression algorithms like *zip*, *gzip*, *bzip2*, are able to recognize the regularities in the data and the *NCD* measure exploits these abilities. Intuitively, strings with similar schemes take up less space when they are compressed together rather than separately compressed. There are not parameters needed to compute the *NCD*, except for the choice of the compression algorithm and its settings. Research conducted from Vitany et al.

shows that the choice of the compression algorithm has a negligible impact in the final analysis. This is facilitated by the use of a massive data normalization in order to minimize the differences in NCD scores calculated with different compression algorithms. NCD is a non-negative number in the range $0 \leq r \leq 1 + \epsilon$ representing the difference between two strings. Obviously, smaller numbers represent greater similarity. The ϵ value in the upper limit arises from imperfection in the compression algorithms, and it is typically less than 0.1. The calculation of the NCD does not require any prior knowledge about the data. The dimensions of the two strings do not necessarily have to be the same. Moreover, in most cases, the results will be better when the strings are of different lengths.

Although it is not possible to compute how close the NCD is from the ideal NID for a pair of bit strings, NCD has reached interesting results. In Cilibrasi et al. [49], it was used to correctly classify data in areas as diverse as genomics, virology, languages, literature, music, handwritten digits, astronomy, and combinations of objects from completely different domains.

Chapter 5

Ensemble Learning

Supervised learning algorithms search, through a hypothesis space, a good model that will make good predictions for a particular problem. Even if the hypothesis space contains features that are very well-suited for a particular problem, it may be very difficult to find a good one. Not all features can indeed discriminate in the same way. Depending on the problem some of them may show an effective discriminative power, whereas other features may not have such power at all. To this aim, Ensemble Learning [51] was proposed as a machine learning approach where multiple learners are trained to solve the same problem. In contrast to standard learning approaches which try to build one learner from training data, ensemble methods try to construct a set of base learners and combine them to improve the results. Base learners are usually generated from training data by means of a base learning algorithm which can be a decision tree, a neural network or other kinds of machine learning algorithms.

Many authors have experimentally demonstrated significant performance improvements through the use of ensemble methods [52, 53, 54]. Boosting is one of the most influential ensemble methods. Its foundation was incite from the answer to a theoretical question posed by Kearns and Valiant about “Probably Approximately Correct - PAC” learning model [55, 56]. They were the first to pose the question of whether a weak learner, which performs just slightly better than random guessing in the “PAC” model, can be “boosted”

into a more accurate strong classifier. Shapire [57] found that the answer to this question is positive, and he gave a proof which is enclosed in the first Boosting algorithm. An important drawback of that algorithm is the requirement that the error bound of the base learners must be known ahead of time, which is usually difficult to achieve. Freund and Schapire [58] then proposed an adaptive Boosting algorithm, called AdaBoost, which does not require this unavailable information.

In the remainder of this chapter we provide a brief introduction to AdaBoost and an illustration of how this algorithm can be adapted to be used in a real application such as the face detection problem.

5.1 AdaBoost

There are many Boosting algorithms. They usually differ depending on their method of weighting training data points and hypotheses. AdaBoost (**A**daptive **B**oosting), is the most popular Boosting algorithm, which was introduced by Freund and Schapire in 1995 [58]. AdaBoost maintains a weighted distribution over the training data and adjusts it at each Boosting round. In particular, higher weights are associated to examples classified with a lower accuracy by the current weak classifier. As a main effect, this weighting scheme forces to focus learning on most difficult examples of training data.

Pseudo-code for AdaBoost is reported in Figure 5.1. The algorithm maintains a set of weights over the m training examples. AdaBoost operates for T Boosting iteration. On each one, a distribution D_t is computed by normalizing the current weights. This distribution is then forwarded to the weak learner which generates a hypothesis h_t . Based on the errors committed by this hypothesis, a new weighted distribution D^{t+1} is generated and the process is repeated. T weighted training sets are generated in sequence and T classifiers are built. A final strong classifier is obtained using a weighted voting scheme: the weight of each classifier depends on its performance on the training set used to build it.

At each Boosting round the parameter α_t measures the importance that is

Given: $(x_1, y_1), \dots, (x_m, y_m)$ where $x_i \in X, y_i \in Y = \{-1, +1\}$

Initialize $D_1(i) = 1/m$.

For $t = 1, \dots, T$:

- Train weak learner using distribution D_t .
- Get weak hypothesis $h_t : X \rightarrow \{-1, +1\}$ with error

$$\epsilon_t = \Pr_{i \sim D_t} [h_t(x_i) \neq y_i].$$

- Choose $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$.
- Update:

$$\begin{aligned} D_{t+1}(i) &= \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases} \\ &= \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t} \end{aligned}$$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right).$$

Figure 5.1: AdaBoost pseudo-code.

assigned to the classifier h_t . The relationship between error and α_t can be described as follows:

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t} \quad (5.1)$$

It means that $\alpha_t \geq 0$ if $\epsilon_t \leq \frac{1}{2}$. Then, α_t gets larger as ϵ_t gets smaller. In other words, if the current classifier is trained for a feature which maintains α_t greater than 0.5 (like the random guessing), then the feature is a good choice in the weak learning algorithm.

Many articles have shown that AdaBoost is often immune from overfitting problem even when a complex hypothesis space is involved, i.e., when T grows. It has also been observed a decrease in the generalization error even when the training error has already reached zero [59, 60].

5.1.1 Real application

Boosting has already been applied in different applications such as text categorization, data mining, object recognition, computer-aided medical diagnosis, and so on. Relevant for the present work is the contribution of Viola and Jones [61], in which the authors propose to use AdaBoost combined with a cascade of strong classifiers for object detection. The Viola-Jones object detection framework provides a real-time detection rate. Although it can be trained to detect a variety of object classes, it was motivated primarily by the problem of face detection. This algorithm consists of three main parts, i.e., integral image calculation, feature definition and extraction, and classification through a cascade of strong classifiers.

Feature extraction

Features used by Viola-Jones are reminiscent of Haar basis functions which have been used by Papageorgiou et al. [62]. The main motivation for the use of these features is that they allow to be easily adapted to recognize different types of objects. Within the face detection context, the authors propose three kinds of Haar-features (Figure 5.2).

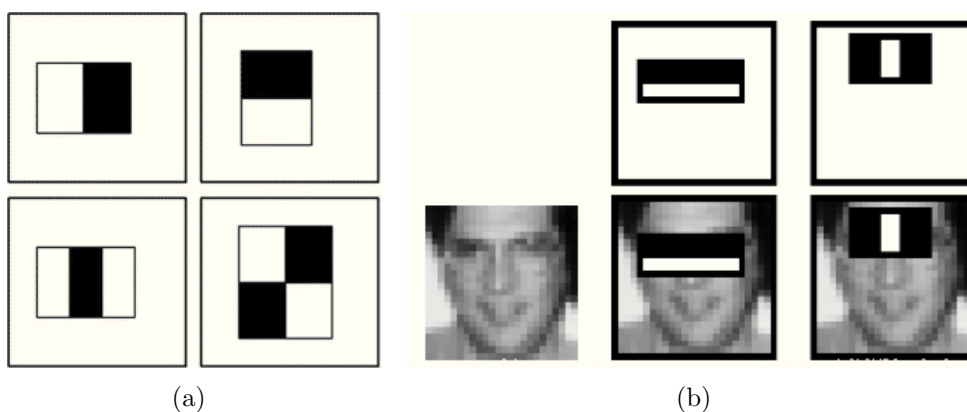


Figure 5.2: (a) Haar features proposed by Viola-Jones. (b) Two relevant features used for face detection. They rely on the intensity contrast between adjacent rectangular areas of the image.

The value of a two-rectangle feature is the difference between the sum of the pixel values within two rectangular regions. The regions have the same size and shape and are horizontally or vertically adjacent. A three-rectangle feature computes the sum within two outside rectangles subtracted from the sum in a center rectangle. Finally a four-rectangle feature computes the difference between diagonal pairs of rectangles. As can be seen from Figure 5.2(b), a simple two-rectangle feature is very efficient to measure the difference in intensity between the region of the eyes and the region across the upper cheeks. Similarly, a three-rectangle feature may help to find the region between the nose and the eyes.

Integral image

Rectangular Haar features used by Viola and Jones can efficiently be computed using an intermediate representation called integral image (also known as a summed-area table [63]). Consider a two-dimensional gray-tone image i . The integral image, denoted $ii(x, y)$, at location (x, y) contains the sum of the pixel value above and to the left of (x, y) , including x and y ,

$$ii(x, y) = \sum_{\substack{x' \leq x \\ y' \leq y}} i(x', y') \quad (5.2)$$

where $i(x, y)$ is the input image. The integral image can be computed in linear time over the image using the following recurrence relation:

$$s(x, y) = s(x, y - 1) + i(x, y) \quad (5.3)$$

$$ii(x, y) = ii(x - 1, y) + s(x, y) \quad (5.4)$$

where $s(x, y)$ is the cumulative row sum, $s(x, -1) = 0$, and $ii(-1, y) = 0$. Given the integral image, the sum of pixel values within a rectangular area of the image can be computed with four array references regardless of the size and location of that area. For example, to compute the sum of pixel intensities inside the region D in Figure 5.3(b), the following four references are required: $L_4 + L_1 - (L_2 + L_3)$. One needs to compute the

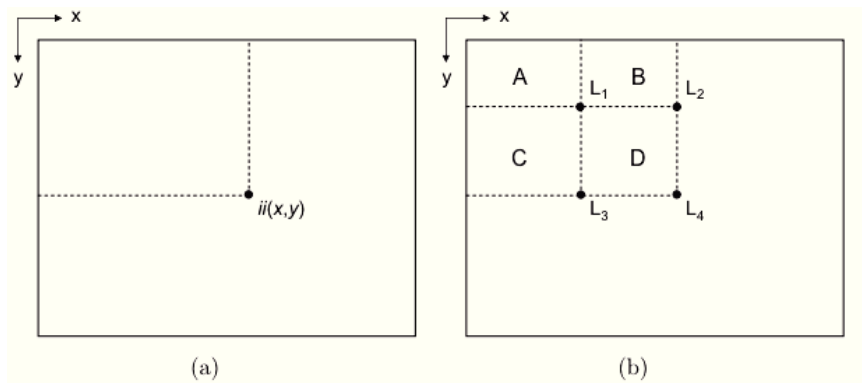


Figure 5.3: Integral image representation. (a) The value of the integral image at point (x, y) is the sum of all the pixels above and to the left of (x, y) . (b) The region D can be computed using the following four array references: $L_4 + L_1 - (L_2 + L_3)$.

integral image only once, and then one can calculate in a very efficient way the pixel sum in any rectangular image region. Since the Viola-Jones two-rectangle features involve adjacent rectangular sums, they can be computed in six array references, eight in the case of the three-rectangle features, and nine for four-rectangle features. Figure 5.4 shows an example of computation of a simple two-rectangle feature.

8	1	2	1	7	16	16	15
9	3	10	4	0	15	15	14
1	15	19	8	8	5	9	12
9	10	8	8	7	15	7	11
6	10	8	5	5	14	13	13
11	9	14	8	6	9	9	13
5	18	9	15	7	10	7	11
12	16	16	15	14	11	8	15

(a)

8	9	11	12	19	35	51	66
17	21	33	38	45	76	107	136
18	37	68	81	96	132	172	213
27	56	95	116	138	189	236	288
33	72	119	145	172	237	297	362
44	92	153	187	220	294	363	441
49	115	185	234	274	358	434	523
61	143	229	293	347	442	526	630

(b)

Figure 5.4: Evaluation of a simple two-rectangle feature. (a) A reference matrix and its integral image representation (b). The light gray rectangle refers to the white region of the feature. To compute feature score six references from integral image are required. $W = 187 + 21 - 38 - 92 = 78$ and $B = 294 + 38 - 76 - 187 = 69$. Feature score $S = W - B = 9$.

Cascade of strong classifiers

Features scores are calculated from sub-windows of reference images. The training module considers images rescaled to a base resolution of 24×24 pixels. Each of the proposed features types are scaled and shifted across all possible combination along an image. This give rise to an overcomplete set of features. Application of AdaBoost provides a list of best discriminative features. In particular, Viola and Jones build a binary classifier for each feature (these are traditionally referred in the boosting community as weak classifiers). Initially all the examples have the same weight. For each boosting step, the determination of a new weak classifier involves the evaluation of the relevance of each feature on training data. The best feature is selected according to the weighted error that each feature shows on the training data. In the successive round, the samples are reweighted to emphasize the misclassified ones. Since this step has to be iterated several times, this is the most expensive section of the training module.

The final step involves the construction of a cascade of strong classifiers. The idea is that simpler classifiers are used at the beginning steps and more complex classifiers are applied at the later steps to reduce the false positive rates. Each node of the cascade is a classifier built through the AdaBoost procedure to satisfy a specific detection and false positive rate. A 24×24 patch is forwarded to the first node of the cascade. If it is judged as a positive example, it is sent to a more complex classifier. A patch is labelled as a face if it successfully overcomes each node of the cascade. A negative outcome at any point give rise to the immediate rejection of the sub-window without further processing. In practice, in a single image, the majority of sub-windows refer to not-face objects. A cascade of strong classifiers rejects most of the negative examples in the initial stages and it will continue to test only in the promising sub-windows.

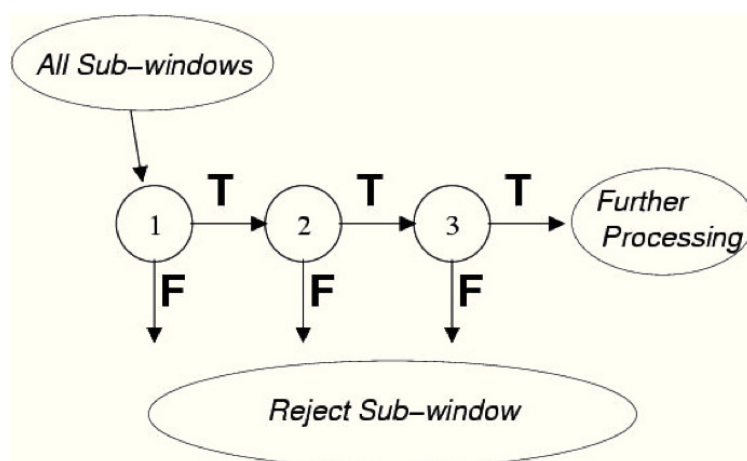


Figure 5.5: Cascade of strong classifiers. A sub-window is subjected to the classification of the first node of the cascade. If it is labeled as a positive example by the current classifier, it is forwarded to a second more complex classifier. A sub-window is labeled as a face if it successfully overcomes each node of the cascade. A negative outcome at any point of the cascade leads to the immediate rejection of the sub-window without further processing.

Chapter 6

Experiments

In this chapter we provide all the results obtained applying the methods that have been described previously. In particular, Section 6.2 reports experiments published in [64], Section 6.3 refers to [65, 66], Section 6.4 refers to [67].

6.1 Dataset

This research has been conducted in collaboration with a group of gastroenterologists from “Maddalena Raimondi” hospital in San Cataldo (Caltanissetta, Sicily). They have provided us the WCE data collected in the period between 2005 and 2010. Each video has been decomposed into individual frames and these have been manually labeled by the expert according to a specific detection task. A typical frame coming from the video-capsule contains a black border in which the information of the patient and the exam are annotated. For all the conducted experiments, we restrict the region of interest within the circular area of the video, hence for each frame only a sub-image is considered (Figure 6.1).

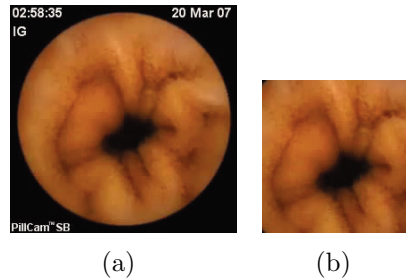


Figure 6.1: Preprocessing of WCE data. Original image (a) and the extracted region of interest (b).

6.2 Information Theory based WCE video summarization

In this experiment an algorithmic information-theoretic method is presented for the automatic summarization of meaningful changes in video sequences extracted from WCE videos. To segment a WCE video into anatomic parts (esophagus, stomach, small intestine, colon), we use a textons-based method [68, 69, 70]. The local textons histogram sequence is used for image representation and the Normalized Compression Distance (NCD) [48] is used to compute the similarity between images.

6.2.1 Feature extraction

In this phase of processing we take into account the *HSI* representation of data. Each frame is also partitioned into sub-squares of 16×16 pixels. For each one of these square sub-windows we extract the features for the next automatic classification. A visual analysis by the clinician is mainly based on a direct examination of the chrominance values of the frames. To this aim, we choose to include the average values of the hue, saturation and intensity of each of the blocks of a frame. These features, although informative, are not sufficient to effectively classify the frames and they should be combined with more information.

As pointed out in Section 3.1, when a capsule travels around a boundary between two different digestive organs, the corresponding color signal has a

sudden change and an increase (or a decrease) of energy. For this reason, we include the high frequency energy content (HFC) of blocks among the features used by the classifier.

Finally, we choose a Gabor filter bank in order to characterize the texture information. In particular, we consider the following parameters set: *phase*: 0, 2, 4, 8, 16, 32 and four directions: 0° , 45° , 90° , 135° .

All the mentioned features have been chosen in order to obtain a good balance between recall and precision of the resulting classifier.

6.2.2 Classification method

Each 16×16 patch is described by a feature vector of 28 components: information on color chrominance from *HSI* color space (3 elements), *HFC* (1 element) and Gabor filter responses (24 elements). In order to achieve a more abstract representation we put together the vectors of all of the 16×16 blocks of the frames in the video. The next step involves the creation of a “bag of visual words” from this ensemble. To this purpose, we use a textons-based method [68]. Feature vectors are initially clustered into *Textons* using a standard K-Means algorithm [71]. The number of clusters is chosen to optimize the ratio of dispersion between cluster centroids over the dispersion within clusters (Figure 6.2). We empirically found that a suitable value for the number of clusters in our experiments is 100. By associating to each patch its reference centroid, this *Textons* dictionary allow us to represent an image as a “bag of visual words”. The histogram of *Textons*, i.e., the frequency with which each texton occurs in an image, represents the model corresponding to each training image. Figure 6.3 summarizes this process.

Comparison between histograms provides a way to assign a distance between consecutive frames in a video. Relevant for this work is the contribution of Gallo et al. [72] in which the Bhattacharya distance [73] between the corresponding histograms of the frames is computed. Bhattacharya distance $d(f_i, f_j)$ of a simple pair of consecutive frames f_i and f_j is generally a weak indicator of changes in the video. This happens because occasionally a frame can be quite different from the previous one just because of casual distur-

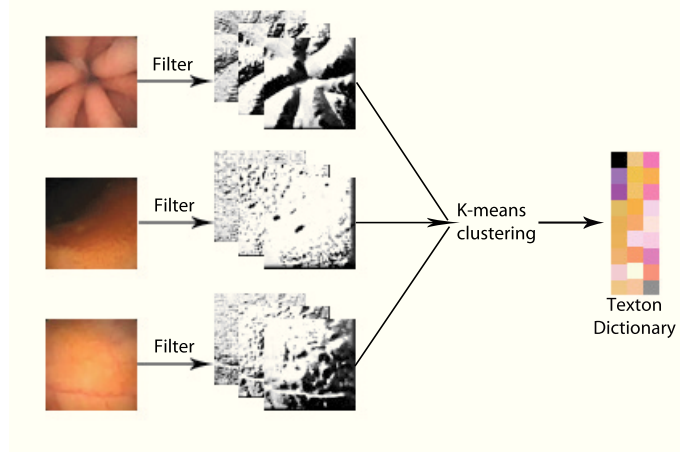


Figure 6.2: A schematic illustration of the *Textons* method. Every image of the training set is convolved with a filter bank. Filter responses are clustered using K- Means algorithm to build a *Textons* dictionary.

bances and transmission noise. To have a more robust indicator of sudden changes in the video, for each frame f_i , an indicator function $C(i)$ is defined to average the distances between frames in a short sequence:

$$C(i) = \frac{1}{9} \sum_{k=i}^{i+2} \sum_{j=3}^5 d(f_k, f_{i+j}) \quad (6.1)$$

The original contribution of this experiment is the use of an information theoretic approach to summarize meaningful changes in WCE image sequences. In our experiments we tested *NCD* distance adopting several compression algorithm (dzip, gzip, etc.) instead of using the Bhattacharya distance, as it has been proposed in [72]. Although these compression algorithms are supposed to grant a good performance because of their ability to exploit the sequential redundancies in the data, their usage is costly. We found that, for the problem at hand, the gain obtained in this way is not relevant and for this reason we introduced a simplified (although rough) version of *NCD* based on Shannon's Entropy:

$$NCD_{entropy}(x, y) = \frac{E(x, y) - \min(E(x), E(y))}{\max(E(x), E(y))} \quad (6.2)$$

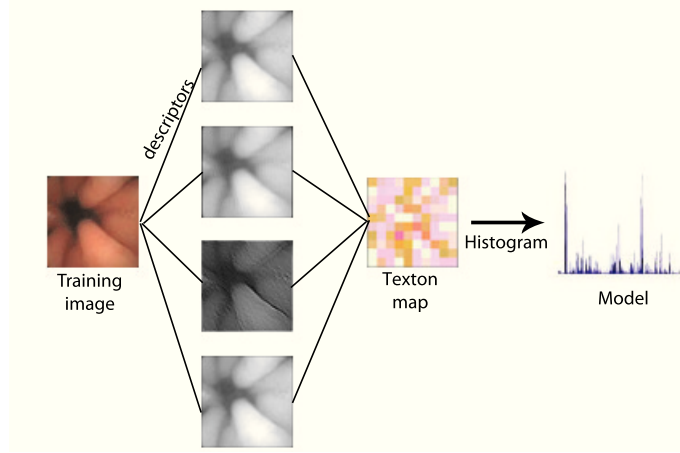


Figure 6.3: Representation of frames as a “bag of visual words”. Each frame is represented by mean of the histograms over the resulting *Textons* dictionary.

where $E(x)$ is the Shannon’s entropy for the string x and $E(x, y)$ is the entropy of the concatenation of the string x and y . In the application considered here x is the string obtained concatenating the “symbols” made with the texton dictionary. In other words a frame from a WCE video is represented here as a sequence of visual words. Following a common practice in Computer Vision, we disregard the sequential order of the words and represent a frame as a “bag of visual words”. This observation justify the substitution of a compression algorithm with the much less expensive use of Shannon’s entropy. The use of entropy in place of Kolmogorov complexity to calculate the *NCD* is not novel even in image domain, see for example [74, 75]. Observe that if sequentiality is disregarded, the entropy of the string of visual words obtained concatenating the representation of two frames is the entropy relative to the averaged histogram of the visual words frequencies in two frames. For the WCE application, however, it makes sense to bias the difference between frames not only considering the visual differences but taking into account the proximity of the frames within the video. To this aim, a new similarity distance *SIM* is introduced as follows:

$$SIM(x, y) = \alpha * NCD_{entropy}(x, y) + \beta * |i(x) - i(y)| \quad (6.3)$$

$i(x)$ and $i(y)$ represent the index of two frames in the video sequence and $\alpha + \beta = 1$. In this experiments the best result have been obtained with $\alpha = 0.8$ and $\beta = 0.2$. Following the approach in [72], for each frame f_i , the expression 6.1 can be written as:

$$Score(i) = \frac{1}{9} \sum_{k=i}^{i+2} \sum_{j=3}^5 SIM(f_k, f_{i+j}) \quad (6.4)$$

$Score(i)$ averages the distances between frames in a short sequence and it provides high values when there is an abrupt change or low values in segments with similar frames. Thresholding the function $Score(i)$ will lead to select interval of frames in which there is a sudden change in pattern.

6.2.3 Experimental results

In this section a number of experiments are undertaken in a real problem domain to demonstrate the efficacy of the proposed method. In our experiments we use ten video sequences provided by the ‘‘Maddalena Raimondi’’ Hospital. We use the labelling protocol explained in [76]. Let $(f_1 \dots f_N)$ be the sequence of frames in a video. We have formed the sequence of intervals $(I_1 \dots I_{\frac{N-3}{3}})$ where interval I_i is made of the six frames $(f_{3i-2} \dots f_{3i+3})$ (Figure 6.4).

In our setting an event includes every change in pattern in a short video

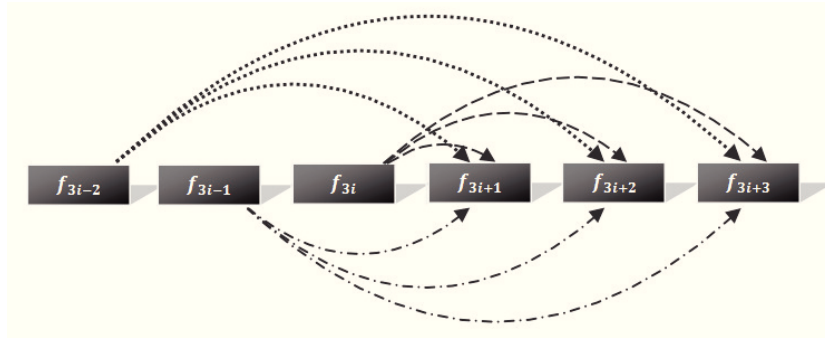


Figure 6.4: The computation of function $Score(i)$. Equation 6.4 is defined to average the similarity distances between frames in a short sequence.

sequence like a boundary transition, a pathology or a common disturbance like intestinal juices, residuals, bubbles, etc. For each interval the clinician has judged if there is a significative change between the first 3 frames with respect to the last 3 frames. If this is the case the interval has been labelled as an “event” (Figure 6.5). To grant greater robustness the labelling has been performed independently by two human experts. Only those intervals that both of them have labelled “event” are considered real event in the following experiments. The two independent labelling agree on 93% of the cases.

Intervals I_i of each video have been sorted according to the decreasing value of their $Score(i)$ indicator. We have hence partitioned the sorted I_i 's into ten groups of the same size. The first group contains the intervals with the top 10% of $Score(i)$, the last group contains the interval with the lowest 10% of $Score(i)$. For each group we have counted the number of intervals labelled as event. In this experimental session two experts have labelled the sequences and the resulting ensemble has been given by their intersection. The bar plot of Figure 6.6 shows the average percent of intervals that have been labelled as event vs the intervals that have been labelled as not-event in the ten groups. The use of precision-recall analysis is investigated in Figure 6.14. As the ROC curve shows the discrimination obtained using the proposed method is comparable with the results in [72]. The slightly less robust discrimination shown by the novel method is justified by the proposed usage of $NCD_{entropy}$ instead of classical NCD . This loss in discrimination power

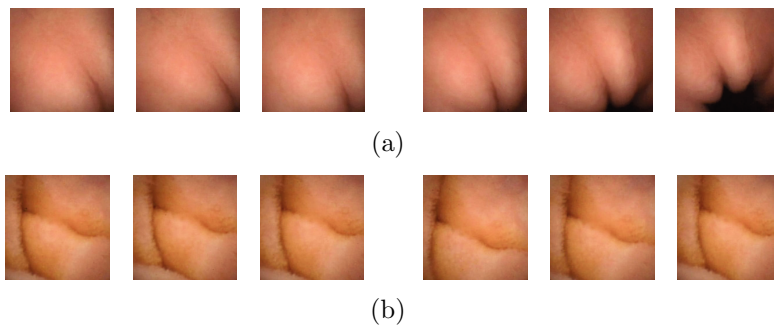


Figure 6.5: Two examples of sequences of consecutive frames. The row (a) represents an event. The row (b) is relative to an homogeneous tract.

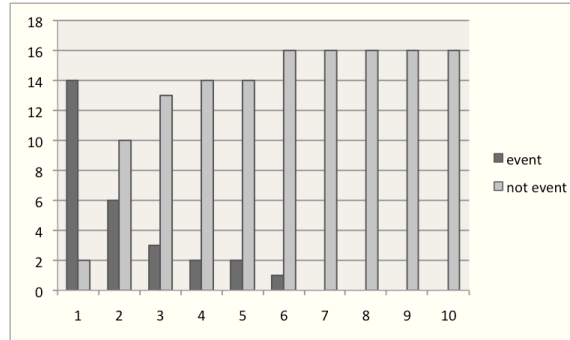


Figure 6.6: Percentage of events and not-events in a WCE video. Intervals of six consecutive frames have been sorted according to the decreasing value of their $Score(i)$ indicator and partitioned in ten group of the same size. The first group (the first column in the graph) contains the intervals with the top 10% of $Score(i)$, the last group (the last column in the graph) contains the interval with the lowest 10% of $Score(i)$.

is however justified by the greater efficiency that the usage of $NCD_{entropy}$ provides with respect to NCD . We also compare the results obtained with the formula of $NCD_{entropy}$ (Equation 6.2), the modified version that uses the concept of entropy, and $NCD_{\alpha\beta}$ (Equation 6.3). Results are shown in Table 6.1.

Some examples of events detected with the proposed method are reported in Figure 6.8. The first row corresponds to an event found in the first 10% until the last row corresponds to an event found in the sixth interval.

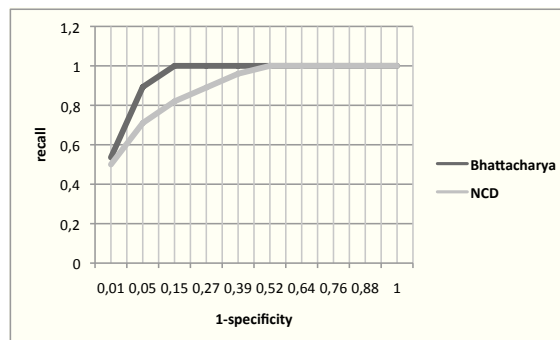


Figure 6.7: Two ROC curves compare the performance of tested methods.

Table 6.1: Summary of experimental results.

Intervals	$NCD_{entropy}$	$NCD_{\alpha\beta}$
top 20%	72%	71%
top 20%	68%	66%
top 30%	85%	86%
top 30%	53%	54%

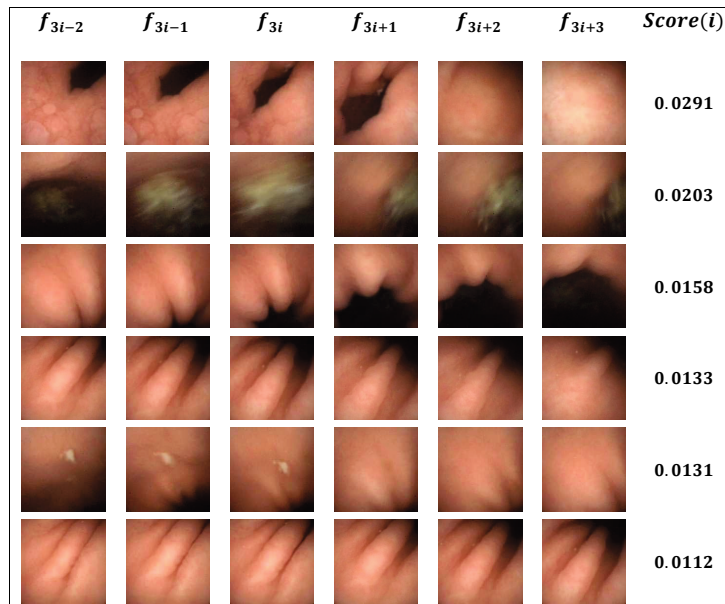


Figure 6.8: Examples of events found with the proposed method.

6.2.4 Conclusion

In this experiment we have presented an algorithmic information-theoretic method applied to find sudden changes in WCE video sequences. We used a modified formula NCD to compute the distance between the histograms obtained with the *Textons* approach explained in [76]. Experimental results have been shown that using the entropy, in combination with two parameters α and β , we reach a recall of 90% with a precision of 52% discarding the 30% of the video. Future works will extend the usage of NCD -like distance since the early stage of *Textons* dictionary construction.

6.3 Lumen Detection in Endoscopic Images: A Boosting Classification Approach

In this experiment we present a novel method to automatically discriminate a relevant subclass of frames. In particular, our classifier sorts the frames in two categories: “lumen frames” (images depicting the stages of an intestinal contraction where the shrinkage of lumen intestine is well visible) and “not lumen frames” (Figure 6.9).

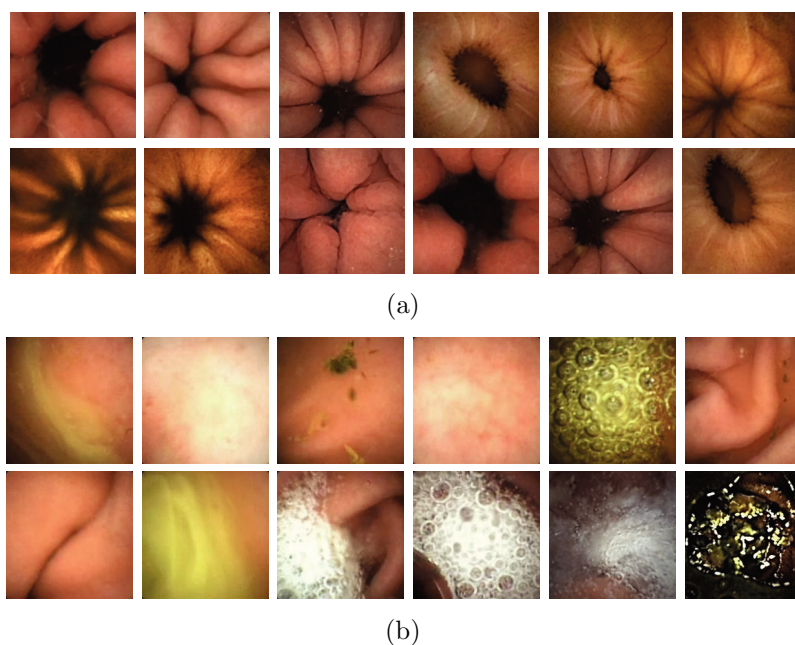


Figure 6.9: Examples of *lumen* (a) and *not lumen* (b) frames extracted from a WCE video.

“Lumen frames” detection is clinically relevant because it announces the presence of a contraction and helps the physician to study the intestinal motility. Alteration of the physiological intestinal motility is an indicator of disorders in which the gut has lost its ability because of endogenous or exogenous causes. In particular, anomalies in contraction are a common symptom of irritable bowel syndrome, delayed gastric emptying, cyclic vomiting syndrome, and so on.

Our summarization tool may be deployed in a diagnostic station providing real-time useful shortcuts to the middle phases of an intestinal contraction resulting in reduced time of analysis by the expert.

In our approach “lumen frames” detection is obtained as a special case of object detection. To this aim, we choose the Viola and Jones paradigm introduced in 2001 [77]. Although other techniques, like neural networks, fuzzy rules systems, etc., could be deployed, the main motivation for our choice has been the following. Haar features based classification is readily customizable to recognize different kinds of objects. Moreover, Boosting allows fast learning even in presence of high dimensionality data. Indeed in the case of Boosting as for all ensemble learning methods, different classifiers are built using a tiny part of the available features. The classification obtained by combining the responses of different classifiers improves the performance achieved by a standard classification algorithm in a straightforward, efficient, principled way when adaptive Boosting is adopted.

In this chapter we describe in detail how the Viola-Jones technique is customized to address the present detection problem. We report the experiment conducted on real WCE videos. To better assess the accuracy of the proposed boosted classifier, we also present an experimental comparison with the results obtained with a Support Vector Machine using a linear kernel.

6.3.1 Feature extraction

The learning stage for the proposed system can be summarized in the following three steps:

- Evaluation of a customized set of Haar features to the integral images of the training samples.
- Selection of the best discriminative features through the application of the AdaBoost algorithm.
- Construction of a final boosted classifier based on a cascade of classifiers whose complexity is gradually increasing.

To obtain, through a reliable learning procedure, a good classifier we must guarantee two requirements: a comprehensive set of examples where the objects of interest may occur; a suitable selection of descriptors to describe each possible occurring pattern. In order to detect an object in an image we should in principle take into account the information provided by each single pixel. This search space may be reduced if we exploit the semantic information enclosed by “lumen frames”. These images, indeed, show a strong geometrical coherence that may help in discriminating them from other kinds of frames. To this aim, Haar-like features, a set derived from Haar wavelets [62], recognize objects using intensity contrast between adjacent regions in an image.

Basic Haar features proposed by Viola-Jones and specialized for face detection do not have proper discriminative power for lumen investigation: it is necessary to define customized variations for the present case. In particular, the features needed in this work should provide a strong positive response on a rectangular region with low intensity called generically “lumen” and a brighter surrounding area corresponding to the gut wall. By combining a learned evaluation threshold to each feature, it is possible to assign an image to the appropriate category. Figure 6.10(a) shows an example of the first kind of our proposed features that we call “center-surround” feature. The typical appearance of a frame that shows an intestinal contraction consists in a dark area surrounded by the typical rays that muscular tone produces due to the folding of the intestinal wall. We hence introduce two additional “cross-like” features that enhance the discriminative power produced by the simpler “center-surround” feature (Figure 6.10(b) - 6.10(c)). The computation of this second kind of features may be efficiently obtained as for the simpler “center-surround” feature from the integral image representation.

Using integral image representation, feature evaluation is accomplished by few memory accesses. It is straightforward to verify that to compute “center-surround” features, at any position or scale, only eight look-ups are needed. The remaining two kinds of features require more accesses due to the greater number of rectangular areas. “Cross-features” require respectively 16 and 24 references from the integral image. The reader may easily convince himself

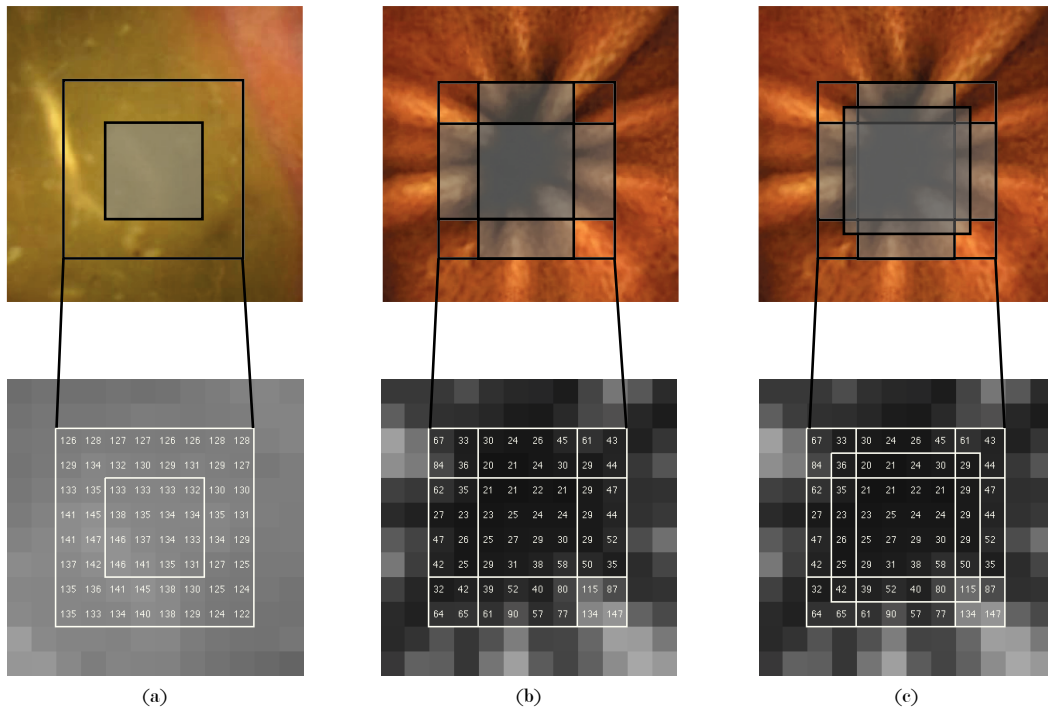


Figure 6.10: The three kinds of features proposed for lumen detection. For each feature we get a score S calculated as the difference of intensity between light and dark regions of the rectangle. In the first row the images at the original resolution are shown while in the second the images are rescaled to the base resolution 24×24 pixels zooming on the region of interest. (a) Evaluation of a “center-surround” feature in a “not lumen frame” ($S_a = 6348 - 2175 = 4173$). (b) Evaluation of the first cross feature in a “lumen frame” ($S_b = 1083 - 1766 = -683$). (c) Evaluation of the second cross feature in a “lumen frame” ($S_c = 861 - 1988 = -1127$).

that indeed this is the minimum number of look-ups needed from a direct analysis of this feature geometry.

Once a feature shape has been assigned, it is necessary to specify its position and scale within the region of interest. Actually, the features are scanned across the image top left to bottom right using a sliding offset of two pixels both in the horizontal and in the vertical directions. The process is iteratively repeated with different feature scales at each round. To keep the computation of the proposed features within the same number of look-ups into the integral image, we choose not to change the scale of the image but

to varyate instead the size of the features. The exact representation for the three proposed types of features is as follows:

$$f = [x_w, y_w, s_{wx}, s_{wy}, x_b, y_b, s_{bx}, s_{by}, type, \theta, \rho] \quad (6.5)$$

The first four elements x_w, y_w, s_{wx}, s_{wy} , refer to the larger square of the feature. Similarly, the following four elements x_b, y_b, s_{bx}, s_{by} , relate to the inner square. The *type* parameter is an integer that indicates which type of feature is considered (1 for the “center-surround” feature, 2 and 3 for the two kinds of cross features respectively). The last two parameters are the optimal learned threshold and the polarity to register the category of images discriminated by that feature.

The “center-surround” features are evaluated considering difference between the sum of the pixels within two rectangular regions (Figure 6.11(a)). The second type of features considers a cross-shaped region to enhance the lumen area. Location and size of this region are constrained by the size of correlated “center-surround” feature (Figure 6.11(b)). The third type of features is processed in a similar way. The central region of the cross is enlarged of one pixel both in the horizontal and in the vertical directions (Figure 6.11(c)). We consider the same total number of features for each type. Lumen area presents always a square aspect ratio, i.e., the bounding region of these areas is approximatively a square. This leads to a simplification of the feature definition as follows:

$$f = [x_w, y_w, s_w, x_b, y_b, s_b, type, \theta, \rho] \quad (6.6)$$

We consider only squared features, i.e., those with equal horizontal and vertical even scale s_w . The internal region relative to lumen varies from a minimum size 2×2 up to $(s_w - 2) \times (s_w - 2)$ pixels. Once we have fixed the size of the external section, the descriptor associated with the lumen is shifted across the external descriptor with a resizing of two pixels at each step (Figure 6.12). In this phase of processing the resolution of a WCE frame is reduced to 24×24 pixels. The total number of features per scale is hence

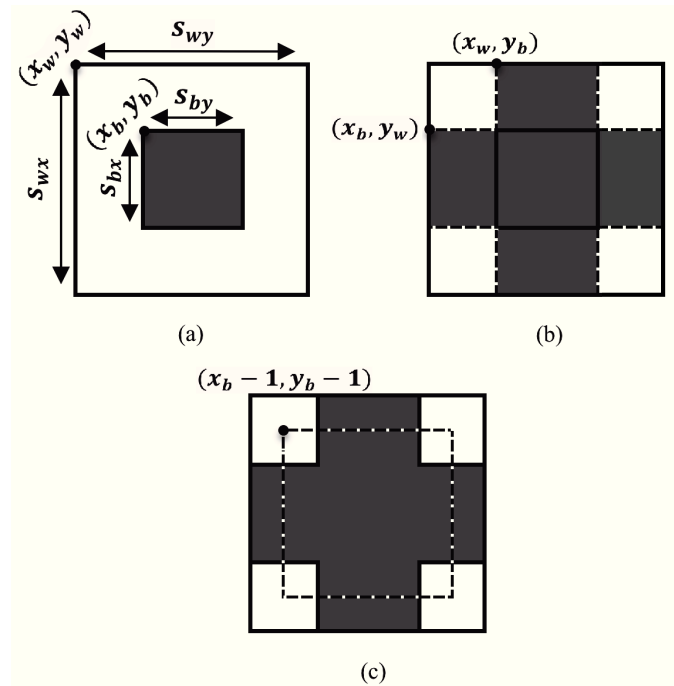


Figure 6.11: Schematic features representation. (a) Center-surround feature. (b) First cross feature obtained by center-surround feature considering the cross with width s_{by} and height s_{bx} . (c) Second cross feature obtained by the first taking into account a inner square of width and height greater than one pixel respect to the previous version.

equal to the total amount of different features in the image multiplied by the allowed variations of scale. For example, a 8×8 feature contains nine regions of size 2×2 , four of size 4×4 and one of size 6×6 pixels. The total number of features of size 8×8 is 1134, equal to the number of windows in the image (assuming a horizontal and vertical offset of two pixels) for the total number of variations. Table 6.2 summarizes the feature counting for the chosen scales.

6.3.2 Classification method

Training a cascade of strong classifiers

As it is stated above, during the training phase the dataset is rescaled to the base resolution 24×24 pixels. The integral image representation of

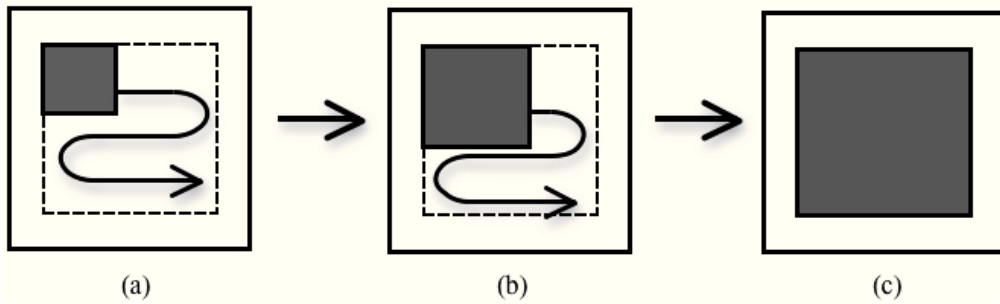


Figure 6.12: Given feature size, all regions of a fixed scale are considered in each location (a). This cycle is reiterated by increasing the size of the inner square (b) until maximum amplitude is achieved (c).

gray tone training samples is used to compute feature scores. Application of AdaBoost provides a list of best discriminative features. In particular, we build a binary classifier for each feature (these are traditionally referred in the Boosting community as weak classifiers). Initially all the examples have the same weight. For each Boosting step, the determination of a new weak classifier involves the evaluation of the relevance of each feature on training data. The “best” feature is selected according to the weighted error that each feature shows on the training data. In the successive round, the samples are reweighed to emphasize the misclassified ones. Since this step has to be iterated several times, this is the most expensive section of the training module.

The result of the training module is a classifier (called “strong classifier” in the Boosting jargon) computed as a weighted linear combination of the weak classifiers built during each round of boosting. The whole Boosting process is, in turn, iterated, varying at each step the number of weak classifiers. The result is the realization of a cascade of strong classifiers with a gradually increasing number of features.

An appropriate learning process requires that each strong classifier shows a prescribed detection rate, while maintaining a definite rate of false positives. In particular, a minimum detection rate and a maximum false positive rate is required at every level of the cascade. For each strong classifier, a weak classifier is added until it reaches the required parameters for the current

Table 6.2: Features number per scale. The first column refers to the size of the feature while the second is related to maximum scale allowed for the lumen area.

<i>Feature size</i>	<i>Max Internal scale</i>	<i>#Features</i>	<i>#Variations</i>	<i>Total</i>
4×4	2×2	121	1	121
6×6	4×4	100	5	500
8×8	6×6	81	14	1134
10×10	8×8	64	30	1920
12×12	10×10	49	55	2695
14×14	12×12	36	91	3276
16×16	14×14	25	140	3500
18×18	16×16	16	204	3264
20×20	18×18	9	285	2565
22×22	20×20	4	385	1540
24×24	22×22	1	506	506
				21021

level of the cascade. Similarly, a new strong classifier is associated to the cascade until total false positive rate crosses a certain threshold.

One of the advantages of the proposed system is that the user only needs to define the feature set to be used and the false positives and detection rates for each level of the cascade. All the internal parameters are automatically selected during the training phase.

Testing a cascade of strong classifiers

In the proposed system, each test image is scaled to 24×24 pixels and it is labelled as “lumen frame” or “not lumen frame”. This single scale procedure combined with selection of best features during training allows real time application of our system (up to 600 frames per second). Please notice that, differently than in the case where the object to recognize may appear at different scales, in the present case a “single-scale” choice has been shown adequate. Notice that in this simplifying choice of a single scale we differ from the original Viola and Jones approach. In the case of face detection

the issue is to find faces that may appear at different scales within an image. These stringent requirements force Viola and Jones to include different scales in their detection procedure. In our case the problem is simpler: lumens are roughly all at the same scale and we do not require localization of them inside the frame but only to label the frame as a “lumen frame”. This justifies our choice of a single scale.

6.3.3 Experimental results

Boosting based classification

In this section, we report the experiments carried out to verify the efficacy of the proposed method. To this aim, we have considered 10033 images extracted from real WCE videos of 12 patients of which 6 were healthy and 6 had suspected bowel disorders. Rather than considering only one training set as was done in an earlier version of this work [65], we have extracted ten different training sets and control sets from the whole set at our disposal. This more extensive experiment has been aimed to verify if the behavior of the algorithm significantly changes according to the used learning set.

To train each one of the cascades of strong classifiers, we take into account the integral images of 3000 images, 1000 positive and 2000 negative, rescaled to 24×24 pixels. The positive images have been previously manually selected from WCE videos labelled by an expert. The selected images represent a comprehensive set of scenes where the intestinal lumen can be present, including location and scale changes within the image. Differently, the negative examples have been randomly selected from videos that not contain any lumen. Both typical smooth images and images containing other judged negative events, like the presence of bubbles, bleedings, residuals, share this set.

During the learning module, we need to establish a maximum false positive rate and a minimum detection rate to satisfy for each layer of cascade. In particular, we require that 98% of positive images must be recognized at each level while maintaining a maximum amount of false positives equivalent to 80%. These values have been experimentally optimized. Notice, however, that higher positive images recognition rate are first of all rarely attainable

Table 6.3: Details on trained cascades using ten different training sets.

<i>Train Data</i>	<i>Nodes</i>	<i>features</i>	<i>Center surround</i>	<i>Cross 1</i>	<i>Cross 2</i>
1	6	217	51	78	88
2	5	291	82	109	100
3	6	397	89	154	154
4	6	342	77	131	134
5	6	256	57	71	128
6	5	185	47	67	71
7	6	257	72	98	87
8	5	205	60	66	79
9	6	184	47	80	57
10	5	272	67	100	105

and even when possible, they may introduce strong overfitting. At the next levels of the cascade these two values are computed relatively to the new dataset whose positives set is composed by every lumen recognized as such by the previous classifier; the negatives set includes the remaining false positives. A strong classifier will be added to the cascade until the total false positive rate drops to zero. By iterating this process for each training set, we get ten different cascades of strong classifiers whose details are listed in Table 6.3. It is straightforward to understand that the trained cascades are slightly different only in the total number of features, but the proportion of features is often the same: the cross-shaped features (*Cross 1*, *Cross 2*) are the most discriminative. The number of nodes in the cascade can not be deterministically calculated, but this also depends on the type of images used during learning. We do not impose any constraints on the number of features in each node. It is assured only that the node $i + 1$ must have a greater or equal number of features than node i . To clarify this procedure, in Figure 6.13 is illustrated the cascade of strong classifiers relative to the 8-th dataset. The total detection rate of this cascade, D , and the final false positive rate F , are obtained as a combination of intermediate outcomes on the cascade:

$$D = \prod_{i=1}^N d_i = 97,98\% \quad F = \prod_{i=1}^N f_i = 0\% \quad (6.7)$$

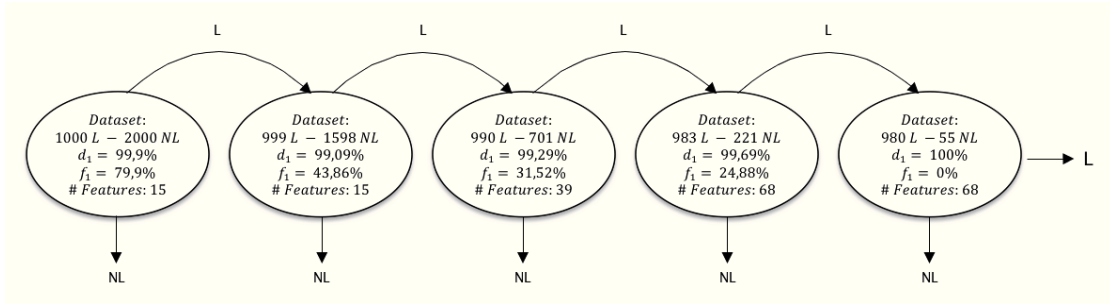


Figure 6.13: Example of a cascade of strong classifiers obtained in the experiments. d_i and f_i represent detection and false positive rate at the i -th level of cascade. L and NL indicate *lumen* and *not lumen* frames, respectively.

where N is the total number of layers of the cascade.

To test the effectiveness of trained cascades, we have considered ten different collections of 7033 images randomly extracted from a set of frames disjointed from each training set. During testing phase, we consider the integral images of test set rescaled to 24×24 pixels with the respective labels, the cascade of boosted classifiers as it has been obtained during training and, finally, a threshold that determines the rigorousness of the classifier. Each test sample gets through each single node of the cascade; a positive outcome is sent by the classifier i to the more complex classifier $i + 1$. An image is labeled as lumen if positively overcomes each node of the cascade. If at any point the test image is judged negative, it is rejected immediately without further test (Figure 6.13). The classification performance has been evaluated in terms of precision and recall by comparing our results with the annotations provided by the specialist. Table 6.4 shows the results. The labeling of images was previously made by a human expert. However, for certain images it is often difficult to understand, even to a skilled human observer, if what we hold as “lumen frame” is actually a particular fold of the intestinal tissue or vice versa.

Each strong classifier in the cascade is constrained by a rigidity threshold. Higher threshold values minimizes both detection and false positive rates. Similarly, a low threshold will lead to acceptance of a greater number of lumens images while increasing the probability of detecting false positives.

Table 6.4: Classification results using Boosting.

<i>Test Data</i>	<i>Recall</i>	<i>Precision</i>	<i>Accuracy</i>
1	88,60%	72,06%	91,32%(6423/7033)
2	89,05%	71,64%	91,24%(6417/7033)
3	91,82%	69,11%	90,67%(6377/7033)
4	91,37%	67,86%	90,16%(6341/7033)
5	87,92%	70,73%	90,81%(6387/7033)
6	88,07%	71,76%	91,17%(6412/7033)
7	88,90%	69,06%	90,34%(6354/7033)
8	90,85%	70,78%	91,15%(6411/7033)
9	86,95%	73,40%	91,55%(6439/7033)
10	91,45%	68,33%	90,34%(6354/7033)

The optimal value of threshold depends on the preferences of the physician. We expect that a higher amount of false positives than of false negatives is typically preferred. The presence of a high number of false positive results in more time spent by the expert to do a diagnosis. Losing a rightful lumen is a worse event because it means to miss a relevant event with the resulting inaccuracy in the final report. By varying the rigidity threshold from a minimum to a maximum value, we can construct a ROC curve comparing the detection rate versus the number of false positives. Figure 6.14 reveals that is possible to reach a detection rate above the 90%, keeping the amount of false positives at about 600 instances, i.e., 8% of the test dataset. All experiments have been conducted on a consumer level PC with Intel®Core™2 Duo processor and 4 GB of RAM. Calculations have been performed in MATLAB environment.

Figure 6.15 shows some examples of false positives obtained with the proposed method. In many circumstances, the intensity contrast between adjacent regions does not correspond to the presence of a lumen. This is maybe a consequence that Haar features are sensitive to illumination changes. Variations on the lighting conditions may cause the cascade to detect lumen that was not predicted during the training stage. Likewise, in some images, folds of the intestinal wall may produce contrasted regions that confuse the Haar

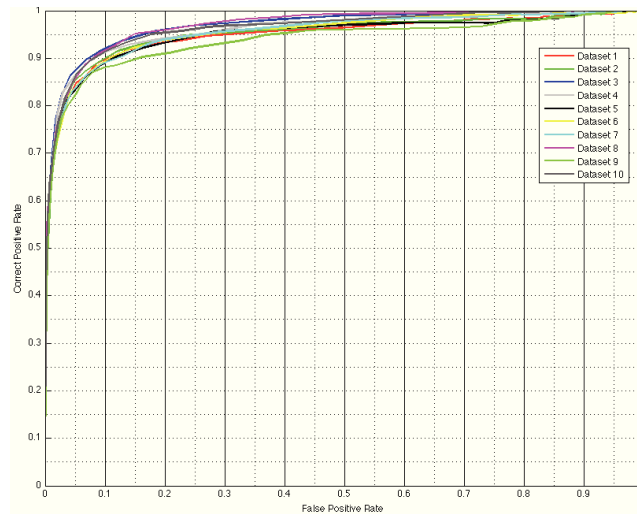


Figure 6.14: ROC curve for each dataset obtained by varying the stiffness threshold of each classifier from 0.1 to 1.

features. If new kind of images are presented to the classifier, detection is difficult and the amount of false positives increases. To deal with this problem, training data must include as many examples as possible to predict only true lumen.

Features analysis

One may reasonably ask if the proposed kind of features is optimal: may we obtain good classification results without one of these three kind of features? May we get away with only one kind? Adding some more elaborate Haar-like features is worth the gain in accuracy? The authors have tried to perform



Figure 6.15: Example of some false positives detected by the system.

boosted classification using only one kind of feature among those proposed in this experiment at each time. The results were only slightly different than those obtained using the whole set of features. This suggests that we might use only one kind of feature and achieve similar results. It is relevant to point out that the cross-shaped features have been introduced by the authors to improve not the results but the stability of the classifier. The availability of the whole set of features helps to keep down the number of classifiers in each node of the cascade. This happens because AdaBoost achieves more quickly the requirements fixed for the current classifier by the user. Also the number of nodes in the cascade is minimized. However, the use of cross-shaped features may bring discriminating power for all those suspicious regions that seem at first sight a lumen, while they are actually residuals, bubbles, or any other intestine artifact that generates an intensity contrast similar to a lumen. We can confirm that the use of additional features can only take effect on the structure of the classifier. The results would not be further significantly improved.

Comparing the boosted classifier with Support Vector Machine

The mean recall value we obtained using boosting is 89,5%. This result is efficiently attainable allowing a real-time performance. An interesting question is to compare the results provided by the boosting-based implementation with another “classic” classification method. The main problem in our data is the excessive dimensionality (63,063 features for each image to be classified). The high number of features suggests that comparison with other classification technique is fair only if these other techniques are adequate to handle these cases. For this reason, Support Vector Machine (SVM) is the ideal candidate for comparison. It is well know that SVM may easily deal with very high feature dimension; moreover, standard SVM implementation are available and this makes comparison easier and repeatable. SVM is a supervised learning algorithm used both for classification and regression. It indicates a binary classifier which projects the training samples in a multidimensional space looking for a separating hyperplane in this space. The

Table 6.5: Classification results using Support Vector Machine.

<i>Test Data</i>	<i>Recall</i>	<i>Precision</i>	<i>Accuracy</i>
1	69, 92%	63, 84%	86, 79%(6104/7033)
2	71, 57%	63, 77%	86, 90%(6112/7033)
3	70, 82%	66, 39%	87, 67%(6166/7033)
4	69, 62%	64, 27%	86, 90%(6112/7033)
5	68, 79%	67, 58%	87, 83%(6177/7033)
6	70, 59%	65, 39%	87, 35%(6143/7033)
7	69, 17%	66, 14%	87, 44%(6150/7033)
8	70, 37%	65, 37%	87, 32%(6141/7033)
9	70, 37%	64, 11%	86, 92%(6113/7033)
10	72, 77%	63, 86%	87, 03%(6121/7033)

hyperplane should maximize the margin, i.e., the distance from the closest training examples. SVM is well adapted to handle the curse of dimensionality and its performance has been tested in different application domains. We have considered the same data used in the previous experiments to train different SVMs using a linear kernel. We rely on a particular class of SVM called Least Squares SVM (LS-SVM). In this version it is possible to maximize the margin between support vectors by solving a linear equation with a least squares method. Classification results using this method are shown in Table 6.5. The superiority of the proposed Boosting based technique is evident.

6.3.4 Conclusion

In this experiment we introduced an automatic lumen detection algorithm for endoscopic images. Inspired by Viola-Jones object detection system, we show that using AdaBoost learning-based algorithm combined with a cascade of strong classifiers leads to a good rate of detection minimizing running time. Experimental results show that the proposed system detects positive images using exclusively Haar-like proposed features. Our detector is flexible and easily extensible to other semantic objects in endoscopic applications.

6.4 Random Forests based WCE frames classification

This experiment provides another methodology for the detection task discussed in the previous Section 6.3. We are satisfied with the performance achieved by the Haar-features set to recognize images containing a clearly narrowing of the intestinal lumen and we repropose their use in this new experiment. Here, the customized set of Haar-like features is used for the growth of different binary decision tree. Each tree assigns a label. One image is eventually associated with the class that has the majority vote in the forest. Experiments conducted on real WCE images have proved the effectiveness of the proposal and are reported and discussed.

6.4.1 Classification method

The proposed automatic lumen detection tool is inspired by the work of L. Breiman [78]. The motivation behind the use of this method comes from its steady success in different classification domains. In addition, the construction of a forest, meant as an ensemble of low level tree-based classifiers, has better performance than individual classifiers and it is more robust to noise. The first stage towards the classification is to extract salient information from crude WCE images. To this aim, we propose the same feature set used in Section 6.3.1 and shown in Figure 6.16.

Feature-score extraction creates a “population” of N records arranged into a matrix D of size $N \times M$. N indicates the number of training samples with M different features. Each entry in this matrix indicates the score of a Haar feature on a sample. This information is used for the recursive growth of a binary decision tree. The process begins from the root node, which takes as input the entire matrix D formulating binary questions that determine whether each record is assigned to the left or the right descendant node. The process is recursively repeated treating each child node as the father of the next iteration until termination conditions are achieved.

We must establish a valid criterion to perform the best split on data. The

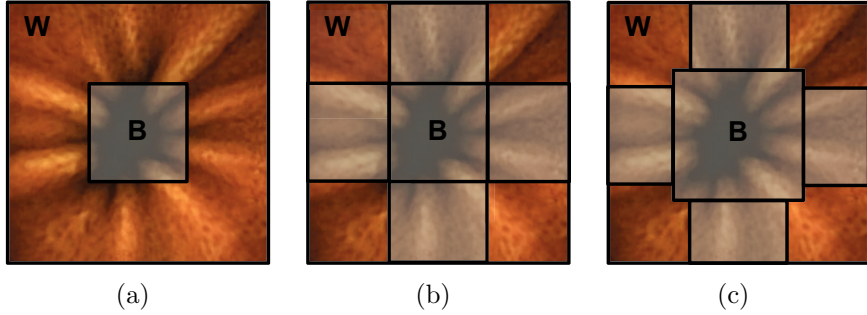


Figure 6.16: The three features used in the proposed method. The score of a feature is computed as: $S(x) = \text{sumrect}(W) - \text{sumrect}(B)$, where sumrect indicates the total value of pixels intensity within a rectangular region. W and B are related to light and dark regions of the feature. (a) “Center-surround” feature. (b) First “Cross” feature. (c) Second “Cross” feature obtained by the previous considering a larger central section.

split function will be good if it helps to achieve a higher degree of homogeneity in each terminal node: the degree of impurities in each node must be minimal. In our experiments Gini coefficient is used to establish the impurity of a node [79]. A node assumes a minimal value zero when the elements in the node all belong to the same class. Gini’s coefficient is defined as follows:

$$g(t) = 1 - \sum_{i=1}^M p(i|t)^2 \quad (6.8)$$

where $p(i|t)$ is the rate of items labelled i in the node t .

A valid criterion for stopping the growth of the tree is relevant for the size of a tree. We consider two alternatives:

- The current node contains elements belonging to a single class: this gives a “pure” leaf node whose label is the label of its element.
- The current node aggregates too few records to allow a new split. The minimal number of records is a user defined parameter and it is useful to avoid overfitting problems. In our implementation we experimentally choose 10 as a good value for this parameter. If an “impure” node is stopped from splitting it is a terminal leaf and it assigns the label of the majority of the records in the node.

This procedure is flexible to process data with either discrete or continuous variables and it does not require any assumption on the statistical distribution of data as in our present application. Once a decision tree has been learnt from training data, classification is very simple and fast, although the classification performance of a single tree may be unsatisfactory.

To improve the classification performance we adopt the *Random Forests* technique, as it has been proposed by L. Breiman in 2001 [78]. The idea is to build many different decision trees. Each of these low-level classifier is parametrized slightly differently than each other. In particular, each tree is grown by combining the technique of Bagging [80] with a random selection of features:

- Let N the number of samples in the training set. For each tree a new training set consisting of N elements randomly chosen (with replacement) from the original data is selected to grow a tree. Small changes in the training set can result in substantial changes in the final result.
- Let M the number of discriminative features for classification. For each split only $m \ll M$ different features are randomly selected. The value of m is set by the user at the beginning of the training phase and remains constant for each split and for each tree.

The mechanism to create a forest from a large training set is very simple: the user needs only to establish the number of different trees in the forest and the number of features to consider for each split. The trees are also not subjected to pruning operation. Observe that each tree uses only a proper subset of the whole training set for its growth. The unused records are indeed useful to establish the classification accuracy of each tree and hence to weight their contribution toward the final label assignment. Given a test sample, it is forwarded to the classification of each tree in the forest. Each tree provides a label. The class with the most votes is the one that is associated with the case. This classification can be carried out in two different ways taking into account that each trained tree has a different accuracy than others. The simplest way of classification is the one that associates a unit weight to each tree in the forest. Otherwise, the final classification may be more

affected by the trees that have greater accuracy. The weight of each tree is computed as the ratio between the number of items correctly labeled and the total number of elements, otherwise known as Out Of Bag Error (OOBE). This index also allows to know which variables are most discriminative for classification, offering the possibility to select a subset that is optimal from the statistical point of view.

6.4.2 Experimental results

We performed our experiments on a training dataset of 3000 images, 1000 of which are “lumen frames” taken at different scales and locations. The remaining images contain other negative events typically seen in a WCE video: from the normal intestinal mucosa to specific events including the presence of bubbles, residuals, bleedings, ulcers, etc. Based on the Haar features scores, we started the training phase for the construction of the forest. As mentioned before, two parameters are required by the user: the number of trees in the forest and the number of different features to be used in each split. Not knowing a priori the optimal number of trees, we trained a forest with a progressively increasing number of trees (up to 100 trees). In each bifurcation, the search for the best split is done by minimizing the Gini coefficient on a subset of $m = \lfloor \sqrt{M} \rfloor$ features¹.

To test the classifier obtained in this way we considered another different dataset of 7033 images. Figure 6.17 shows recall and precision as trees are added to the forest. All experiments have been conducted on a consumer level PC with Intel®Core™2 Duo processor and 4 GB of RAM. Calculations have been performed in MATLAB environment.

As can be seen from the picture, the performance of the classifier does not improve after a certain number of trees have been added. Experiments have shown that 50 is a safe choice for the optimal number of trees. Table 6.6 shows the classification results obtained with this choice of parameters and a comparison with the related boosted-based technique [65]. Random For-

¹The number 21021 in Table 6.2 refers to the total number of features for type. We consider the same total number of features for each type, for a total amount of $M = 63063$ features.

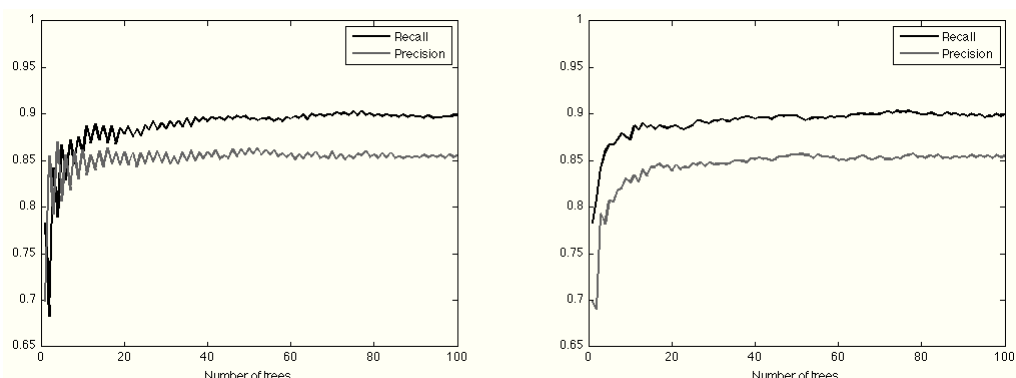


Figure 6.17: Comparison of recall and precision rate as a function of the number of trees in the forest. The graph on the left refers to the testing phase in which each tree has a unit weight. The graph on the right takes into account the OOB error.

est finds a slightly smaller amount of lumens, but the false positives rate is greatly minimized compared to the boosting based implementation. Also the accuracy is greatly improved.

Notice that our technique allows the exploration of the discriminative power of each feature. We used this opportunity to study which of the three proposed types of features is most discriminative for the “lumen problem”. Surprisingly, the “center-surround” feature provides less detailed information. This feature takes an internal area with low intensity surrounded by a lighter landscape. No one indeed ensures that the internal region refers to a lumen. The same circumstances can occur in other types of images. For example, there are images with low intensity residuals, they also surrounded by a lighter background (see the third image in Figure 6.9(b)). The second and third type of proposed features enrich the discriminatory power and thus are more relevant to the classification (Figure 6.18).

Finally, we report some typical errors of our classifier. As can be seen from Figure 6.19(a), many false positives are due to intensity contrast that does not corresponds to any lumens but confuses the behavior of Haar features. False negatives (Figure 6.19(b)) often contain a lumen with a severe offset from the center of the image. In order to resolve these problem, new features should reflect these particular scenarios. In addition, training data must

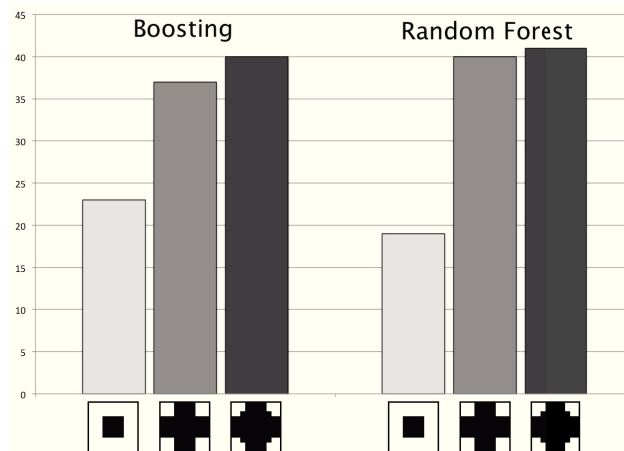


Figure 6.18: Percentage distribution of the three types of features in the final classifier. Either Boosting and Random Forests testify the most discriminatory power provided by the “cross-shape” features.

include as many images as possible to predict only true “lumen frames”.

6.4.3 Conclusion

In this experiment we have reported a classification method to discriminate “lumen frames” in a WCE context. Relying on a custom set of Haar features combined with a classification technique based on randomized trees, we achieved good classification results. Although we referred to images coming from video capsule, our experiments can be reposed for other kind of endoscopic sequences. Our ongoing work involve the strengthening of the proposed algorithms using temporal coherence in a sequence and the generalization of the proposed technique to the classification of other relevant kind of events in endoscopic images.

Table 6.6: Classification Results

	<i>Boosting</i>	<i>Random Forest</i>
<i>Recall</i>	92, 2%	89, 8%
<i>Precision</i>	67, 1%	85, 6%
<i>Accuracy</i>	89, 9%	95, 2%
	(6326/7033)	(6696/7033)

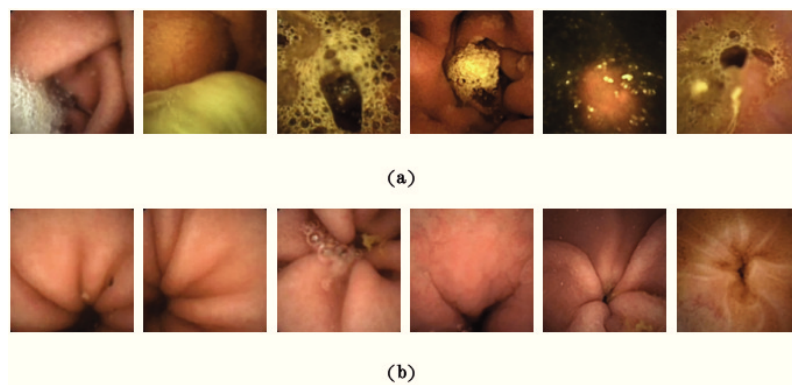


Figure 6.19: Some misclassified of our classifier, false positives (a) and false negatives (b) respectively.

Chapter 7

Conclusion and future work

In the first part of this dissertation we have presented computer-based methods to tackle the problem of automatic classification of endoscopic frames. In particular, the set of images used in our experiments has been obtained by means of the WCE procedure, an endoscopic minimally-invasive technique. As with all the endoscopic examinations, the medical expert needs to see the entire set of images recorded during the exam to do a confident diagnosis. In an effort to keep low the analysis time by the expert, we have conducted research on two different areas: “sudden changes discrimination” and “intestinal motility detection” in a WCE video. In both cases, it is expected to substantially reduce the number of images to be manually analyzed allowing a more widespread use of WCE.

One of the main problems addressed in WCE is the segmentation of the video into homogeneous sections with the same semantic content. In our setting, we have indicated with “event” an abrupt and significative change in the video. It can be represented by a boundary transition from an organ to another one or other relevant events, like the presence of intestinal juices, bubbles, ulcers, etc. To automatically detect such events, we have constructed an indicator function that reveals a sudden change in a video. The construction of the function uses the statistical *Textons* approach combined with an algorithmic information-theoretic method applied to find sudden changes in WCE video sequences. In particular, we have used a modified formula of

Normalized Compression Distance (NCD) to compute the distance between the histograms obtained with the *Textons* approach. The best results have been achieved considering a combination of features related to colours, texture and energy information. The experiments have been demonstrated that the proposed method may eliminate up to 70% of the frames from further processing while retaining all the clinically relevant ones.

The second problem that we have tackled refers to the automatic searching, in a WCE sequence, of the central frames of an intestinal contraction. By labeling these frames, also called “lumen frames”, the physician can easily obtain the video subsequences representing the intestinal contractions and then conduct the related analysis of the intestinal motility. A “lumen frame” consists of a central area with low intensity and a brighter surrounding area corresponding to the gut wall. To recognize this kind of frames, we have proposed a set of image descriptors based on intensity contrast between adjacent regions in an image. The classification task has been performed with Ensemble Learning techniques. In a first experiment we have built a classifier using AdaBoost, a subclass of Boosting, i.e., a machine learning algorithm for performing supervised learning. AdaBoost generates a new “weak classifier” in each of a series of rounds. For each iteration, a distribution of weights is updated that indicates the relevance of each example in the training data. The weights of each incorrectly classified example are increased and the new classifier will focus on the examples which have so far eluded correct classification. This procedure leads to the construction of a “strong classifier” meant as a set of classifiers constructed during each Boosting round. The final step towards the classification involves the creation of a cascade of “strong classifiers” with progressively increasing complexity. It has been proven the effectiveness of ensemble techniques in such systems and showed how a cascade of strong classifiers provide both high accuracy and few false positives for an efficient approach to find intestinal contractions.

The same classification problem has been addressed using a different ensemble technique that uses a set of decision trees instead of AdaBoost as base learners. Each decision tree in the forest is built to train only a tiny part of the available data. Also the hypotheses space is reduced by considering

only a random subset of them. The results obtained using this technique are comparable with those obtained through the AdaBoost procedure. Within the lumen detection task, the use of *Random Forests* approach is able to reduce the false positive rate more than using Boosting. However, if we want to set a stiffness threshold in the final classifier to restrict or enlarge the rate of detected “lumen frames”, we suggest the use of the Boosting approach. In this case, it is possible to conduct a ROC analysis just by ranging the values for the threshold in each weak classifier. The same procedure with the *Random Forests* technique requires the modification of the thresholds for each branch of the decision tree.

The classification domains discussed in this dissertation represent only a portion of what can be automatically classified and recognized in the WCE context. It should be noted that the search for an event can often be rough. In other words, it should be more useful to obtain a list of events divided by type. Recognizing other kind of events (such as bleedings, cancer, polyps, etc.) in a video sequence will help the physician to reduce the time inspection and to make capsule endoscopy a clinical routine. It would be ideal to have a software tool that automatically processes the video content and produce a final report, replacing the work of the expert. Although such a solution is not yet available, there is no doubt that the Computer Vision research on WCE videos will become an important field of medical Image Processing and will gain much wider interest of the researchers in the coming years.

Finally, we are in a continuous feedback with the experts in order to improve the current methods, create optimal protocols and include faster and more efficient versions of our solutions for their use in a real clinical scenario in a close future.

Part II

Depth estimation in Bronchoscopic Intervention

Chapter 8

Stereoscopic Vision

In Computer Vision, 3D reconstruction refers to the process used to obtain a three-dimensional description of the observed scene. There are different reconstruction techniques based on different principles and each one with specific strengths, limitations and areas of application. Among these, the Stereoscopic Vision is the one that has received the most attention.

The idea behind Stereoscopic Vision consists of a triangulation targeted to relate the projection of a point of the scene in two (or more) image planes of the cameras that compose the stereo system. The identification of these homologous points allows to obtain a quantity called disparity by which, knowing the appropriate parameters of the stereoscopic system, it is possible to deduce the 3D location of the considered point. This process is an imitation of one of the capacities of the human vision system where this task is automatically performed by the brain. It merges the retinal images that come from the eyes and perceives them as a single image. It is this process, known as Stereopsis, which allows the relative location of visual objects in depth, giving the perception of the three-dimensional space.

In this chapter we focus our attention on binocular stereo vision with two cameras, called left and right camera, that observe the scene from two different points of view, as illustrated in Figure 8.1. In the first part of this chapter, we introduce the most important parameters of a stereoscopic system required to define the transformation of a three-dimensional point into a

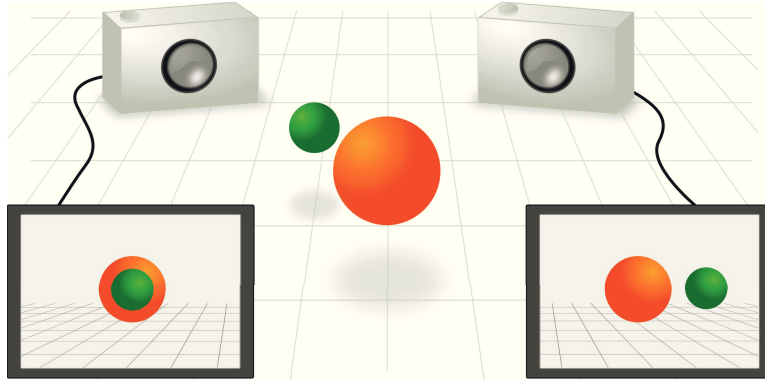


Figure 8.1: Example of a binocular stereo system. The cameras observe the scene from two different point of view.

two-dimensional point of the camera image plane. In the second part, we give some informations about the stereo correspondence problem, i.e., the process by which two projections of a point in two image planes are associated. We illustrate how the epipolar geometry may help to reduce the research space for the correspondence problem.

8.1 Stereoscopic system

8.1.1 Disparity

Consider the stereo system in Figure 8.2. This system is composed by two cameras C_l and C_r with focal points O_l and O_r and parallel optical axes α_l and α_r . The cameras share the same focal length f and T is the horizontal baseline between them. c_l and c_r are the principal points, i.e., the intersection points between the image planes and the optical axes of the cameras. A generic point P is located to a depth Z from the baseline. It projects the points p_l and p_r in the left and right image planes of the cameras, respectively. The pixel p_r in the right image appears shifted to the left compared to the pixel p_l in the left image. This shift between the coordinates x_l and x_r of homologous pixels is called disparity:

$$d = x_r - x_l \quad (8.1)$$

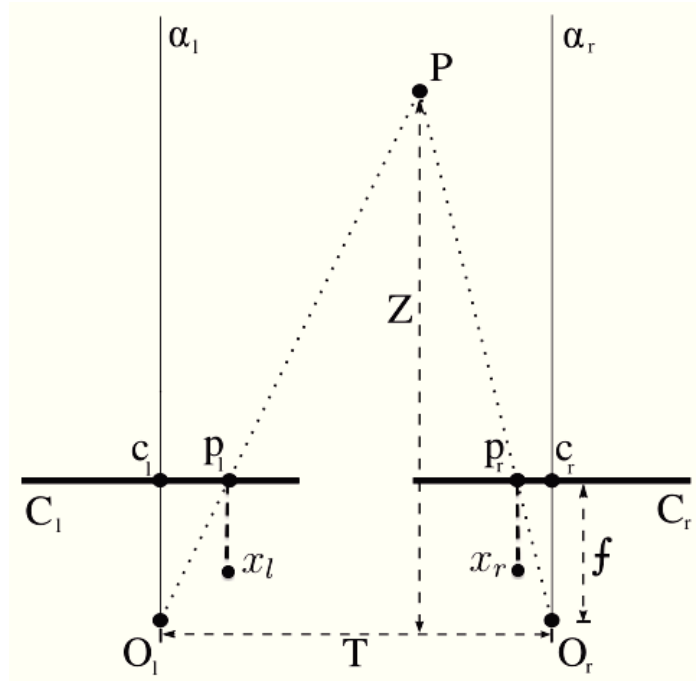


Figure 8.2: Scheme of a binocular stereo system with parallel optical axes. A point P generates two different projections in the image planes of the cameras. The difference between the horizontal coordinates of these projections is called disparity.

Taking into account the similar triangle $p_l\widehat{P}p_r$ and $O_l\widehat{P}O_r$, we can write the following:

$$\begin{aligned} \frac{T}{Z} &= \frac{T-d}{Z-f} \Rightarrow \\ (Z-f)T &= Z(T-d) \Rightarrow \\ TZ - Tf &= TZ - Zd \Rightarrow \\ Z &= \frac{Tf}{d} \end{aligned} \tag{8.2}$$

Equation 8.2 allows to assert the following:

- The disparity value is inversely proportional to the distance separating the point P from the camera planes, i.e., the depth of P . Moreover, depending on the parameters of the stereo setup, the maximum possible disparity between the left and right images can be established.

- The relation between Z and d is not linear.
- Once fixed the focal length and the baseline, the depth of a point depends only on its disparity value.
- Estimation errors of the disparity lead to large errors in the estimated depth.

8.2 The correspondence problem

The main issue in Stereoscopic Vision is represented by the correspondence problem: finding the pixels in two images corresponding to the same point in the scene. This process may be very difficult due to the presence of many complex objects in the scene, often with very similar textures. The situation is worsened by the presence of noise, occlusions or other artifacts that make difficult the coupling of homologous points. A correct match is required for a reliable depth estimation. Consider the stereo system in Figure 8.3. The points P and Q project four points in the image planes of the cameras C_l and C_r . If p_l and q_l are correctly associated with p_r and p_r respectively, the reconstruction will report the depth values of P and Q . If the point p_l is wrongly associated with the point q_r , it will be generated the disparity value for a point P' , totally wrong respect to the real value. The same happens when q_l is matched with p_r , resulting to the depth value of the point Q' . There are two different approaches to solve the matching problem. The “dense” techniques try to find corresponding points between different images maximizing a measure of similarity in the pixel domain. Another approach generates disparity values only in correspondence of a specific set of salient points of the image.

Dense stereo matching methods

Dense matching methods, also called area-based methods, estimate the disparity of a pixel p in the first image (chosen as reference image) comparing a small region around p with all the possible areas of the same size in the

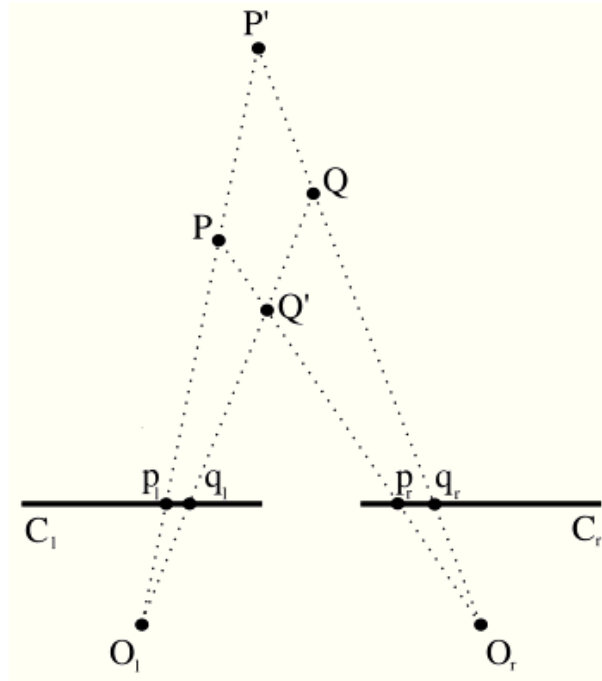
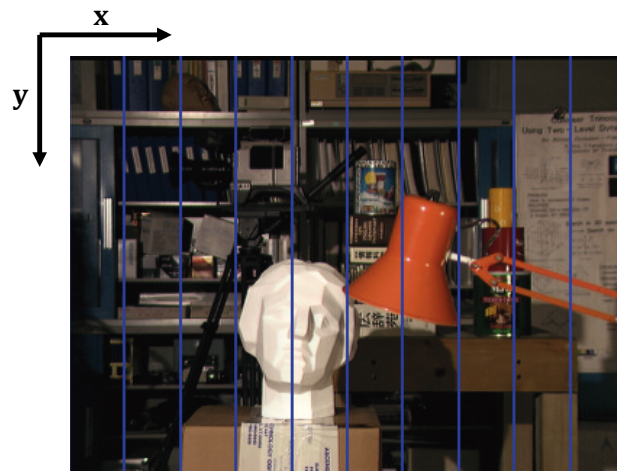


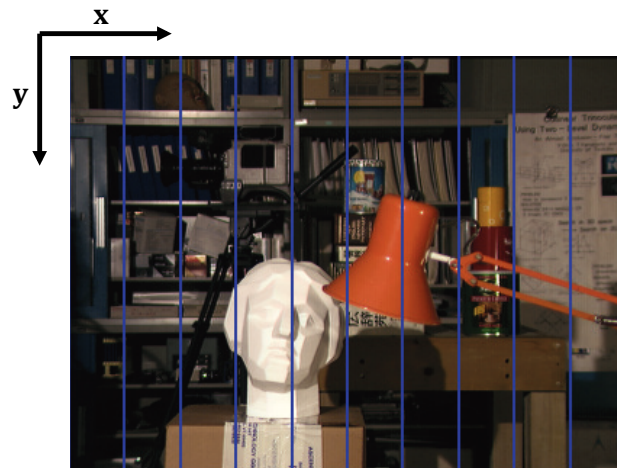
Figure 8.3: If the projections of the points P and Q in the left and right image planes are not correctly associated, it is impossible to estimate the correct depth for these points. This occurs when p_l is wrongly associated to q_r , generating the disparity value for a point P' . The same happens when q_l is matched with p_r , resulting to the depth value of the point Q' .

other image. Once the window is identified as the most similar to the one in question, a correspondence between projections is achieved. The correct disparity can be defined as the difference (in absolute value) between the horizontal coordinates of p and p' , i.e., the centers of the correlation windows whereby the matching function presents the highest peak. A very good metric to measure the similarity between two pixel areas is the statistic correlation. For each pixel in the reference image, a match in the other image can be established. Thus, the correlation leads to the construction of a dense disparity map. If we interpret such map as an intensity image according to the disparity, we will see the areas closest to the camera colored white and distant objects in a color darker and darker.

In Figure 8.4 is illustrated a standard stereo pair. In this example, left image is chosen as the reference image. Notice that the right image is translated



(a)



(b)

Figure 8.4: An example of a standard stereo pair. An object in the left image (a) is shifted across the right respect to the right image (b).

towards the left in relation to the reference image. This shift is due to the different positions of the optics of the cameras. An object in the right image is slightly shifted to the left respect to its representation in the left image. This means that a point P in the left image with coordinates (x, y) has a conjugate in the right image with coordinates $(x - d, y)$, where d is the value of disparity for the point P . It ranges from 0 and a specific maximum value which depends on the hardware setting.

To calculate the similarity between two pixel areas the following measures

are usually adopted:

Sum of Squared Difference

$$C_{SSD}(x, y, d) = \min \left\{ \sum_{i,j} [I_l(x+i, y+j) - I_r(x-d+i, y+j)]^2 \right\} \quad (8.3)$$

Sum of Absolute Difference

$$C_{SAD}(x, y, d) = \min \left\{ \sum_{i,j} |I_l(x+i, y+j) - I_r(x-d+i, y+j)| \right\} \quad (8.4)$$

Normalized Cross Correlation

$$C_{NCC}(x, y, d) = \max \left\{ \frac{\sum_{i,j} [I_l(x+i, y+j) - \bar{I}_l(x, y)] \times [I_r(x-d+i, y+j) - \bar{I}_r(x+d, y)]}{\sqrt{\sum_{i,j} [I_l(x+i, y+j) - \bar{I}_l(x, y)]^2 \times [I_r(x-d+i, y+j) - \bar{I}_r(x+d, y)]^2}} \right\} \quad (8.5)$$

The index i ranges from $-[N/2]$ to $+[N/2]$ where N is the width of the correlation sub-window. Similarly, the index j ranges from $-[M/2]$ to $+[M/2]$ where M is the height of the correlation sub-window. $I_l(x, y)$ and $I_r(x, y)$ denote the intensity values in the point (x, y) for left and right images respectively. Finally, $\bar{I}(x, y)$ is the average intensity value inside the $N \times M$ sub-window. The estimation of the disparity for a point can be obtained through the following three steps:

- Comparing the sub-window around the current point with all the corresponding sub-windows that fall within the range of disparities in the second image.
- The comparison is done using one of the functions defined above over the gray tone images.
- The disparity of the point is obtained from the difference of the horizontal coordinates of the centers of the sub-windows that are the most similar according to the correlation function used for the correspondence.

Performing the comparison between all possible pairs of sub-windows is extremely expensive. Although the corresponding areas have similar coordinates, the number of comparison to be carried out is prohibitive for a real-time system. However, by exploiting a particular feature of the geometry of a stereoscopic system, it is possible to restrict the research of homologous points in one-dimensional space. This will be clearer when we will introduce the epipolar geometry.

Features-based stereo matching methods

A feature is a significant characteristic of the image. There are specific algorithms for detecting different types of features (edges, contours, corners, SIFT, etc.). The first step to compute the disparity involves the use of a feature detector in order to extract a set of features from both left and right images. Each detected feature provides a descriptor. Next, a match function is used to establish the correct correspondences and the disparity can be measured by the coordinates of homologous points. Feature based techniques, although presenting improved accuracy around the extracted features, usually produce sparse disparity maps and require image interpolation to achieve a denser representation. The number of disparities is indeed proportional to the amount of features properly coupled by the matching function.

8.3 Epipolar geometry

The epipolar geometry helps to simplify the matching problem: given a point projected in the image of a camera, the epipolar geometry can limit the search for the homologous point to a single line in the other image. Consider the stereo system in Figure 8.5. There are two cameras C_l and C_r with focal points O_l and O_r . Let p_l and p_r the projections of the point P in the cameras C_l and C_r respectively. Suppose that the cameras have convergent optical axes such that the projection of each focal point intersects the visual plane of the other camera. These points e_l and e_r are called epipoles. The point P and the focal points O_l and O_r form a plane called epipolar plane.

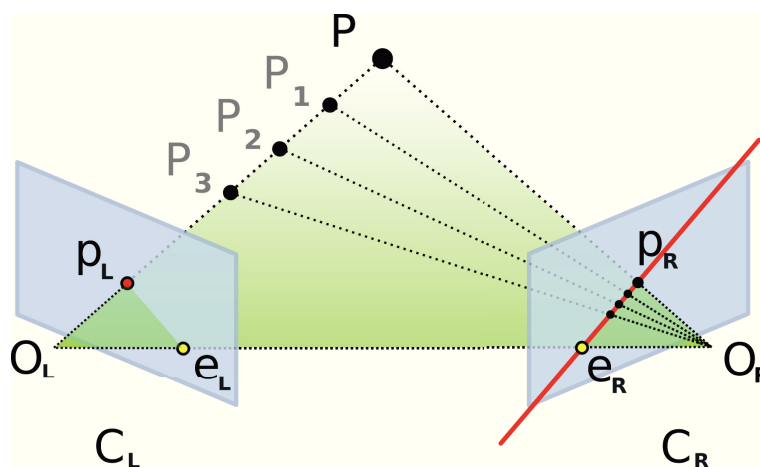


Figure 8.5: Graphical representation of the epipolar line (red line) associated with the point p_l . p_r must be one of the points P_i of this line.

This intersects each camera's image plane in two lines, called epipolar lines. The intuition behind the epipolar geometry is that, if we know the extrinsic parameters of the stereo system and the position of p_l on the image plane of C_l , we can look for p_r only on the corresponding epipolar line on the image plane of C_r .

8.3.1 Calibration

A Stereoscopic Vision system is characterized by the intrinsic and extrinsic parameters. The first ones characterize the camera and enable to model the projection of a point of the scene on the image plane of the camera. These parameters include the coordinates relative to the image plane of the principal point, the focal length, and other parameters that describe further characteristics of the sensor, like the lens distortion and the shape of the pixels. The extrinsic parameters represent the positions of each camera with respect to a known reference system. The determination of the intrinsic and extrinsic parameters, obtained by the calibration procedure, allows to completely describe the stereoscopic system and in particular to infer information about the coordinates of points in space through the triangulation of homologous points.

A further step to simplify the search for corresponding points is the rectification of the cameras' projections. If the optical axes of the cameras are not parallel, the epipolar lines of each camera converge on the epipoles. However, if the epipolar lines are parallel and perfectly horizontal, the search for correspondence may be reduced to a one-dimensional problem (along a horizontal line). Given a pair of stereo images, the intrinsic parameters of each camera, and the extrinsic parameters of the stereo system, the rectification computes the image transformation that makes epipolar lines collinear and parallel to the horizontal axis. In other words, it converts a general stereo configuration to a simple stereo configuration with parallel optical axes.

Chapter 9

Literature Review

In endoscopy, the surgeon can investigate the inside of a body structure (abdomen, intestine, lung, etc) without any surgery, but using an endoscopic camera located at the end of a medical probe. Modern technologies have led to the development of endoscopes that include plenty of accessories and utilities. At the same time the resolution of the integrated optics has been significantly improved. Despite these technological innovations in modern endoscopic surgery, the visualization that is currently used remains 2-dimensional. This is associated with significant limitations, such as the lack of depth perception. It is often quite difficult to locate the objects that can be seen through the endoscope and understand what is deeper in presence of multiple objects. In [81] it has been demonstrated that severe errors made during surgery procedures are not due to poor technical skills but rather reflect a critical misinterpretation of the video image.

The entire endoscopic examination is performed by watching the video monitor while the expert handles the endoscope from outside the body. When the displayed scene is not immediately recognized, or when the image is rotated with respect to the surgeon's perspective, the surgeon often becomes disoriented. This is especially true in the bronchoscopic context where the presence of several ramifications in the bronchial tree makes easy to lose track of the surgical probe.

To compensate these shortcomings, systems that provide 3D visual informa-

tion have found an increasing number of medical applications. A 3D medical system, that is capable of providing more depth clues, has a great potential to improve the current 2D imaging technologies.

While most researches seek for a hardware solution to build a 3D system, there are not many works trying to apply stereovision techniques to extract 3D information from bronchoscopic video. The first experiments of stereovision in medicine also deal with the study of the benefits of stereo viewing rather than the reconstruction which can be started from it. For this reason, in the first part of this chapter we introduce the main technologies currently used in medicine to obtain a 3-dimensional view. Then, we discuss some reconstruction techniques with the aim of using the Augmented Reality.

9.1 Stereoscopy in medicine

The effectiveness and usefulness of stereo property has been suggested in many medical applications. Today, there are two main technologies available that create stereoscopic images in the medical context. The first one involves the use of a surgical probe equipped with two cameras. It captures two pictures that are displayed to the viewer on a stereoscopic display. This kind of technology is adopted in the integrated system “DaVinci” [82], a surgical robot which permits the fulfillment of different surgical operations. It provides an immersive operating environment for the surgeon by providing both high quality stereo visualization and an interface that directly connects the surgeons hands to the motion of surgical tool inside the patient’s body (Figure 9.1). The main disadvantages that exist with dual-camera technology are related to user side-effects such as fatigue, headache, dizziness, and eye strain, that result from viewing two images that differ in picture angle, brightness, color, optical distortion, and sharpness.

A newer different technology, developed by VisionSense [83], uses a single sensor composed by many micro-lenses looking at different directions. The layout of the lenses is similar to the eye-structure of a bug. This technology has the advantage of generating an image from a single charge coupled device, to avoid the problem of dual-camera technologies. The use of a mi-



Figure 9.1: The surgical robot “DaVinci”. It consists of four entities: 1) a plenty of surgical instruments and 2) the console to handle them by the surgeon (3). His assistant (4) works next to the robotic arms and the patient.

croarray of lenses creates multiple small images that are then divided into simultaneous left and right images using proprietary software. The viewer’s eyes then simultaneously pick up two slightly different images of the same object. This provides the surgeon with real-time, high-resolution, natural stereoscopic vision (Figure 9.2).

The introduction of stereoscopy in the medical field has also led to the study of the benefits of this technology with respect to the 2D standard approach. For instance, the authors in [84] report a comparison between the 2D and 3D technologies applied during laparoscopic surgery. By utilizing the “DaVinci” robot launched in both 2D and 3D modes, they evaluate the outcomes of suturing and knot-tying tasks completed by seven participating surgeons. The overall set of experiments was completed 65% faster using 3D mode with equal, if not greater, accuracy. Despite there are limitations to this study design, notably the restricted number of participating surgeons, it is believed that the 3-dimensional view enhances the performance with regard to speed, accuracy, and ease with which different surgical interventions are performed. If stereoscopic viewing includes significant improvements in the diagnostic practice, many improvements also come from the reconstruction obtained using data extracted from a stereo system. The recovery of 3D information

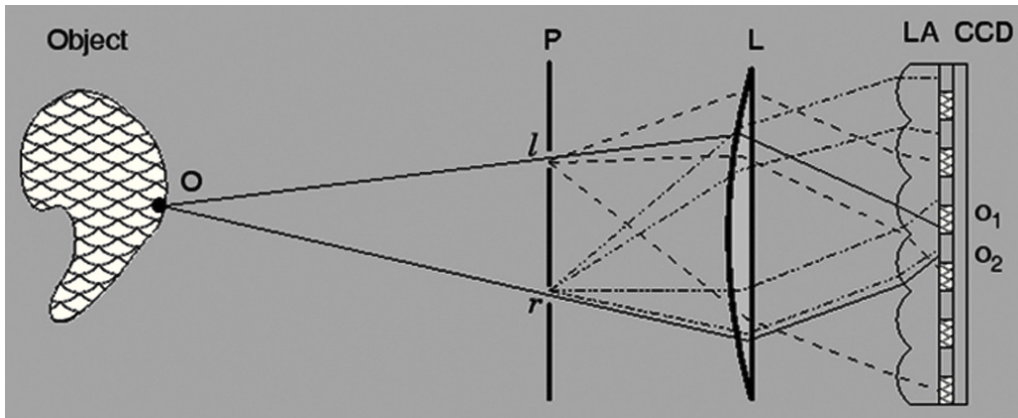


Figure 9.2: Schematic representation of the VisionSense technology. The imaging device is represented by a single lens (L) with two pupils (l and r) located to the focal plane P . All the rays passing from a point O through the pupils generate signal on both pixels O_1 and O_2 . These pixels are left and right views of point O .

from stereo images is one of the classic problems in Computer Vision. A comprehensive summary of progress in this field is reported in [85]. However, most of the reported methods are indicated for regular scene. These contain lots of shape features such as corners, edges, and it is easy to locate the features and find correspondences between adjacent views. Endoscopic images are often characterized by the presence of small field of view, big distortion, varying illumination and surgical instruments. These problems make the reconstruction procedure very difficult. A number of stereoscopic techniques for recovering 3D shape have been proposed in different medical procedures [86, 87, 88, 89]. These experiments are usually performed using endoscopic sequences from the “DaVinci” surgical system.

9.2 Augmented Reality in surgery

The lack of depth perception during endoscopy often limits delicate dissection or suturing [90]. The use of an Augmented Reality interface allows to combine supplementary imagery as part of the scene and can be used for guidance, training and locational aids. An Augmented Reality display presents

objects in correct perspective depth, assuming that the geometry has been accurately acquired. With such a system, a surgeon is able to understand better the structure of the environment by the presence of “artificial” informations that would be impossible to infer from the original visualization.

The authors in [86] propose an Augmented Reality tool to assist the surgeon during cardiac surgery. In particular, they build a time-variant 3D model of the beating heart using standard coronarography sequences, MRI or CT-scan data of the heart. In this experiment, the authors rely on the use of the surgical robot “DaVinci”. In this way, the tridimensional data of the cardiac surface is easily achieved from the stereoscopic optics positioned at the extremity of the surgical probe. Once the observed surface is registered with the 3D heart model, informations about the location of coronary arteries are superimposed on the endoscopic images by Augmented Reality.

It is worth mentioning the work in [91], not for the reconstruction module but rather for the application of Augmented Reality applied in the bronchoscopic context. During a bronchoscopy, the expert inserts a bronchoscope into the bronchial tree using the nasal or the oral cavity. Because of the very complex tree structure of the respiratory system, he/she may be easily confused and loses the way to the target location. For this reason, the authors in [91] propose an augmented display of anatomical names of bronchial branches on real bronchoscopic views in order to improve the navigation and perform the bronchoscopy in safety. To generate such Augmented Reality interface, they initially extract the bronchial tree structure from 3D CT images. For each detected branch, they collect five features: the length of the branch, the running direction of a branch, the relative position from the parent branch, the average direction of the child branches, and the running direction of sibling branches. Then, the proposed method calculates candidate branch names and groups them depending on which branch is the parent. Finally, multi-class AdaBoost technique is used to train the classifiers. To overlay anatomical names on real bronchoscopic images, the proposed method detect the branch where the bronchoscope is currently located and the child branches of the current branch. To find the location of the bronchoscope the authors adopt the Deguchi’s method [92].

Chapter 10

Experiments

The recovery of 3D structure during a medical endoscopic procedure is a necessary step towards accurate deployment of surgical guidance and control techniques. Taking in account the computational stereo theory, the first step to achieve a reliable 3D structure of the scene involves the estimation of disparity, i.e., the apparent pixel difference or motion between a pair of stereo images. A disparity map contains the apparent motion in pixel for every point and it is represented as an intensity image out of these measurements. Large disparities are encoded as light gray values, small disparities as dark grey values. Although it may seem a limited description of a 3D model, it is possible to encode the gray tones present in a disparity map with the actual depth values by using the intrinsic parameters of the cameras.

We believe that depth information may be exploited to reconstruct the scene and to add virtual objects and information to the real images by Augmented Reality. In order to understand the strategy followed to validate this proposal, we report in Figure 10.1 the three kinds of images examined in the experiments. The analysis begins by taking into account real bronchoscopic images (Figure 10.1(a)). In this way it is possible to give a first look at the informative content of this type of images (e.g. texture, lighting conditions) as well as the presence of noise or occlusion, and then figure out which methods should be taken into account in the following stereo analysis. In this phase of investigation it is also possible to verify which characteristics may create

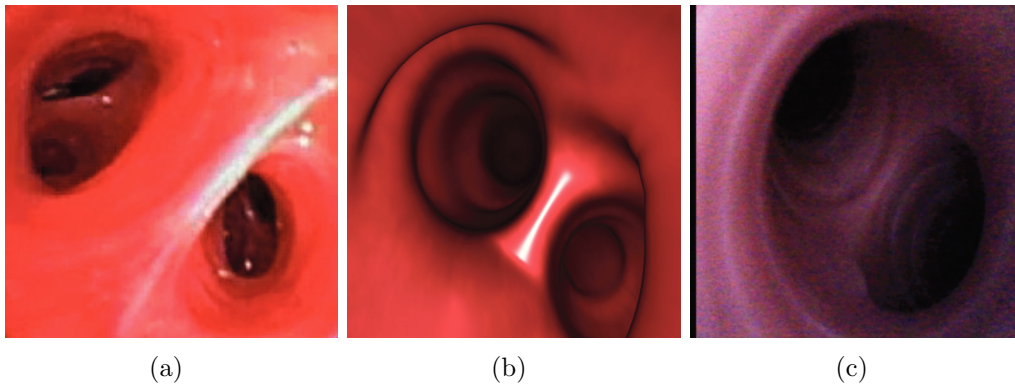


Figure 10.1: The three type of bronchoscopic images used in the experiments. (a) A real bronchoscopic image. (b) A “virtual” bronchoscopic image taken inside a graphic model of the bronchial tree. (c) A bronchoscopic image captured inside the bronchial tree of a medical simulation dummy.

problems in this type of images. Two relevant worsening of images quality are indeed due to non-uniform illumination and blurring. The illumination is not the same in all the images but strongly depends on the distance between the light of the camera and the shooting object. When the camera is too close to a bronchial item, some parts of the images are more illuminated, generating color-saturated areas, while others present more shadows. Blurring effect appears when the camera moves too fast and also when the camera objective is obscured because of the breath of the patient. However, the major drawback encountered during this first analysis is based on the lack of stereo data. In other words, the images on which we work are taken from a monoscopic bronchoscope. To this aim, the second step in our analysis involves the use of virtual reality for a simulation of a stereo bronchoscopy. Having analyzed the anatomy of the respiratory tract, it was possible to create a graphical model of the bronchial tree and navigate inside it with a pair of stereo cameras. An example of frame captured inside the model is reported in Figure 10.1(b). All the problems that can occur in a hypothetical real-world application are here minimized. The use of virtual reality also allows the possibility to perform different experiments with different parameters. For example, it is possible to find the optimal baseline between the cameras just by conducting several navigations inside the model with different baseline values and choose the

one that produces the desired stereo effect. Although the virtual images are not realistic as in the previous case, they contain adequate detail to validate our proposal about the extraction of depth clues for Augmented Reality purposes.

The final step in our analysis involves the creation of a real prototype of a flexible stereo bronchoscope. Figure 10.1(c) reports an example of image captured by one lens of this prototype inside the bronchial tree of a medical simulation dummy. The use of a real stereoscopic instrumentation represents an interesting case. In this way, it is possible to analyze more realistic images compared to the experiment in virtual environment. We can also deal with the problems that might occur in a hypothetical real-world application, notably the lens distortion, lighting control, the focus of the lens.

In summary, the three investigation phases (there are reported in the next three sections) are the followings:

- Depth extraction from monocular bronchoscopy (Section 10.1): in this first experiment, we extract depth information from a standard bronchoscopic examination. By applying the stereo theory to pairs of adjacent images in the video, we extract the disparity in proximity of salient points of the image.
- 3D reconstruction in virtual reality (Section 10.2): disparity maps are obtained this time using a conventional stereo system. In particular, a virtual model of the bronchial tree is crossed by a couple of aligned cameras with parallel optical axes. The convenience of the virtual environment to set optimal values for the cameras have led to good results that have been published in [93].
- Stereoscopic bronchoscope prototype (Section 10.3): in this last work, which also is our ongoing activity, we propose a first prototype of a flexible stereo bronchoscope. After giving enough details about the hardware setting and the software needed to manage it, we reported the same experiments set of the previous work performed on images captured inside the bronchial tree of a medical simulation dummy.

Depth information may be exploited to reconstruct the scene and to add virtual objects and information to the real image by Augmented Reality. For this reason, in the last two experiments reported above, we propose an Augmented Reality application based on the gray tones of the disparity map. It is only one of the possible Augmented Reality interface that can be used in this type of images. It shows the benefits produced by Augmented Reality in bronchoscopy, and more generally in the endoscopic field.

10.1 Depth extraction from monocular bronchoscopy

In order to determine the 3D position of a physical point in space, at least two images that capture the scene from two different points of view are required. The movie obtained by a moving bronchoscope may provide the images of the operating site from distinct viewpoints. Therefore, if the surface does not significantly change over time and if the camera does not change orientation too quickly between two adjacent frames, the 3D information can be extracted by applying computational stereo theory.

In this experiment we take into account a monocular video of a bronchoscopy. Unlike most other systems that utilize stereo endoscopes with dual lens, we study the feasibility of 3D reconstruction from single-lens bronchoscopic video by extracting disparity information around a set of candidate feature matches. Having the availability of an adequate number of disparities, disparity maps have been constructed from two consecutively collected images on the video. The approach presented here consists of two parts. First we estimate the camera parameters based on conventional feature point detection and correspondence analysis. Then, we use the camera motion information to bring two adjacent frames of the video in a standard stereo form. For all pairs of images for which this process is properly performed, disparities are extracted for a set of candidate feature matches. In order to have a denser representation, we use a region growing algorithm to expand the information of disparity on a point to the homogeneous region around it.

The purpose of this experiment is not reduced to the construction of a depth map of the scene. Before reaching this result, it must be analyzed the informative content of this type of images that should be taken into account from the Image Processing point of view. This analysis allows to understand how much detail contain the textures and hence the optimal parameters to be used during the feature extraction step as well as the threshold to employ during the segmentation phase.

The approach used in this experiment to obtain a representation of the depth of the scene highly depends on the movements of the bronchoscope during

the navigation. Note that the movement that occurs between each pair of consecutive frames is not always the same through the video. Even when the movement is minimal but enough to provide depth clues, the number of matches established between two images may be insufficient to ensure a reliable result. For example, this happens when the pair of frames under consideration presents highly blurred areas in which it is difficult to establish any match. However, we expect that there are pairs of frames that simulate a real good stereo pair and at the same time contain a high number of correct matches. For these pairs of frames we expect a reliable disparity map.

10.1.1 Depth clues extraction

The first step towards depth information extraction involves an epipolar rectification step to transform the images in a new version where epipolar lines are parallel along the horizontal plane and corresponding points have the same vertical coordinates. Lacking the availability of the intrinsic parameters of the camera, the only way to estimate the camera motion information relies on the knowledge of some correspondences between two different views. To this aim, we use the Matlab toolbox available at [94]. This framework provide non-calibrated stereo using Fundamental Matrix only, i.e., a 3×3 matrix F of rank 2 that encapsulates the intrinsic projective geometry between two views. If a point in space X is projected in x in the first view, and x' in the second, then the image points satisfy the relation $x'Fx = 0$. The framework finds matching points between two images using the SIFT algorithm [95] and the final rectification is achieved by a suitable rotation of both image planes according to the epipolar constraints. Due to misalignments, usually some of the correspondences are incorrect. To achieve a robust estimation of camera motion parameters, a random sampling algorithm [96] for outlier detection is employed. The reader is referred to [97] for more details.

Once one has obtained a standard stereo pair, our approach finds again the SIFT features from both images and puts them in correspondence with the provided descriptors. In particular we use the Lowe's matching function [95], in which the ratio of the distance of a given feature from its nearest

match to distance from the second nearest match is adopted as a metric. For each correspondence the disparity is calculated as the difference, in absolute value, between the horizontal coordinates of the matching points. Having the availability of rectified images, it is easy to remove from the disparity computation those points for which the vertical coordinates are very different because probably there is the presence of false matches.

When disparities for all matching points of an image have been extracted, a “semi-dense” disparity map of the observed scene can be created. One possible strategy is the use of a segmentation technique for partitioning the image into homogeneous regions. With the availability of appropriately segmented images, the cluster membership of each disparity point can be established. The disparity value for the cluster is represented by the average of the disparities which lie in the same cluster. To segment the images we rely on the use of a region growing approach. This algorithm involves the choice of initial seeds. For each of them, a cluster is created and iteratively grown by comparing all unallocated neighbouring pixels to the cluster. The difference between the value of intensity of a pixel with respect to the average intensity already present in the cluster is used as a criterion of similarity. The pixel with the smallest difference is allocated to the region. In addition to the choice of initial seeds, a threshold is chosen to make more flexible or less the addition of pixels to a cluster. In other words, the segmentation for a given seed stops when the intensity difference between the region mean and a new pixel becomes larger than the threshold. Figure 10.2 shows some examples of the application of the region growing technique with different values of threshold. By conducting the appropriate experiments, it was noticed that it is not easy to establish an optimal threshold value. In this kind of images, a small threshold is suitable to recognize the so-called “blob”, i.e., the entry of a new bronchial branch that appears in depth in the image. Similarly, a too high threshold creates too large regions that make rough the calculation of the disparity with consequent loss of information. In this experiment a threshold equal to 0.07 has been chosen as optimal value.

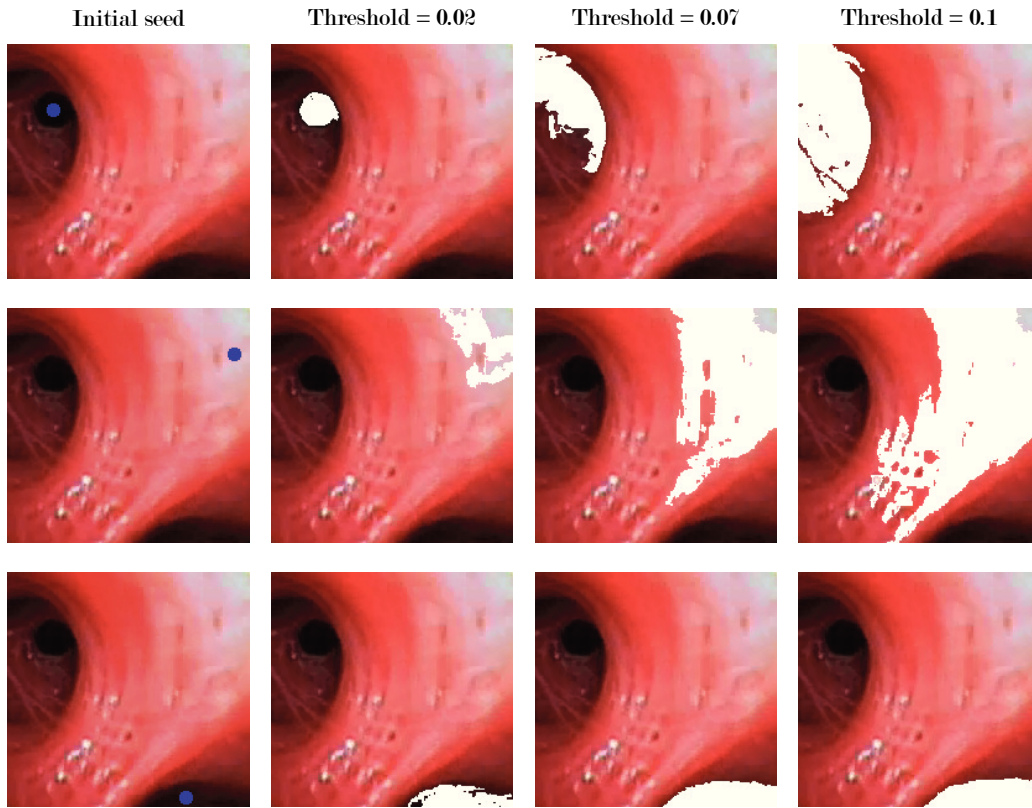


Figure 10.2: Example of application of the region growing technique. The blue point is the initial seed from which a cluster is created and iteratively expanded. The segmentation for a given seed terminates when the difference between the intensity average value inside a cluster and the intensity of a new non-allocated pixel becomes larger than the threshold.

10.1.2 Experimental results

In this section we report the experiments carried out on a monocular bronchoscopic video. Each pair of adjacent images in the video is undergone to an epipolar rectification step and disparity is extracted from the matching points detected on the rectified images. The greater the number of matches, the better the description on the depth resulting from the disparity map with the assumption that there are no matching errors. Figure 10.3 shows some examples of disparity maps obtained with this approach. The “Sparse disparity” column reports the disparity value in correspondence of matching

points. Taking into account the gray tones representation, the white color is associated with the points that display large values of disparity. The greater the depth of the point, the darker the gray tone with which it is represented. For each point we consider all the disparities that fall on the same region according to the region growing application and the final value of the disparity is the average of the disparities of the points that fall in the segmented region (the column “Semi-dense disparity” in Figure 10.3). To ensure that the segmented region does not have rigid edges, we have slightly smoothed the image with a median filter prior to application of the region growing algorithm. Notice that it is not possible to have depth information for all the regions where there are not matching points (the black pixels in our results). For this reason we use the “semi-dense” nomenclature in our final results.

10.1.3 Discussion

One of the major drawbacks in this proposal is certainly due to the lack of stereo information arising from a real pair of stereo cameras. In our experiments we use the motion information between two adjacent views of a monocular video to rectify images so that they can simulate a real stereo pair. However, it may not suite for the whole video sequence since the change of orientations of rectified stereo pairs may not be smooth, which implies unstable results. Moreover, the “virtual” baseline of rectified stereo pairs may also not be the same throughout the video. Some pairs of images are so different that the procedure for rectification can not be properly operated. Disparity maps for those couples can not be achieved.

Another consideration regards the application of region growing segmentation to fill a denser disparity map. For each matching point the algorithm considers it as initial seed. When a pair of images has a high number of matches, the performance of this segmentation algorithm becomes computationally prohibitive. A further relevant problem concerns the choice of the threshold. This strongly depends on the current image and a a-priori estimation is quite difficult. The threshold value adopted in our experiments is not necessarily the optimal one throughout the video.

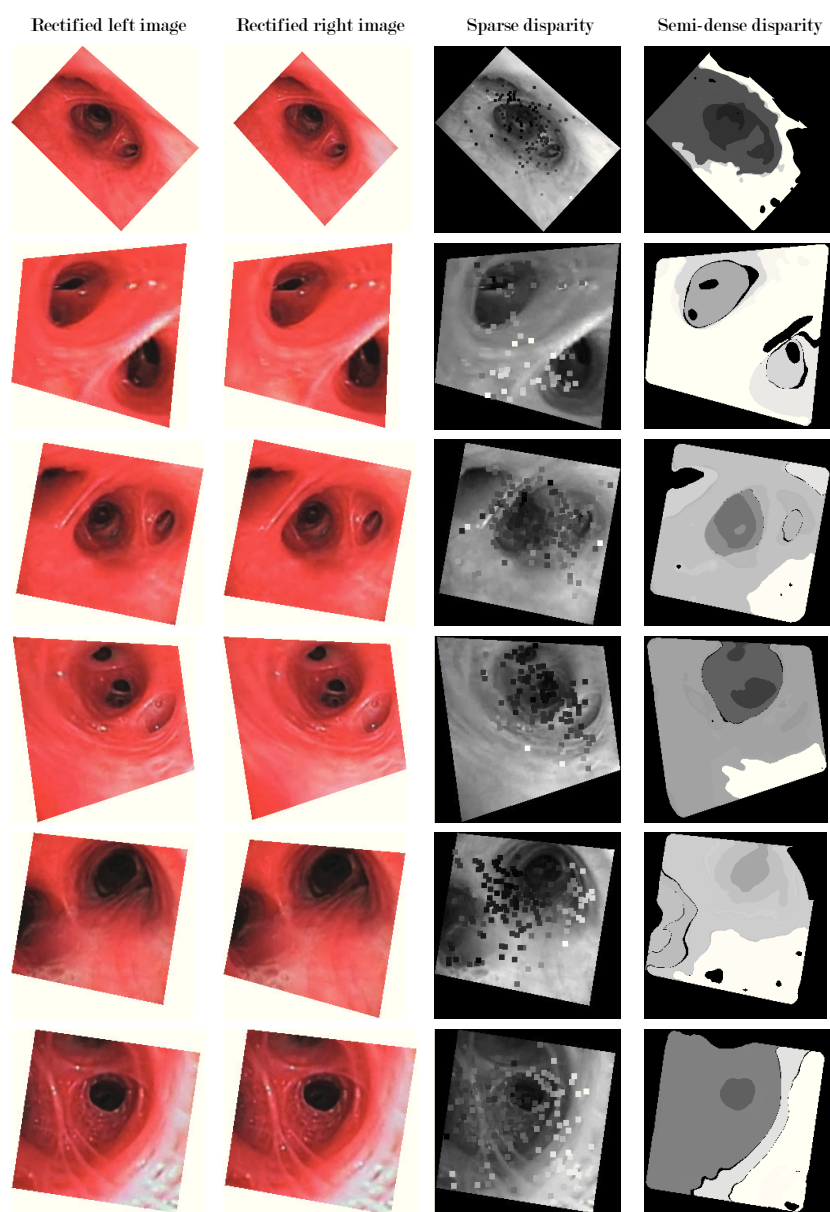


Figure 10.3: Disparity maps obtained using a monocular bronchoscopic video. Each pair of adjacent images in the video is undergone to a rectification step. The consequent extraction of key-points from these images with a suitable matching function leads to the creation of a coarse disparity map. The use of a segmentation technique allows to expand the value of disparity in a point to the whole homogeneous region around it.

Despite these drawbacks, it is easy to verify the validity of our assumptions for all those pairs of images capable of simulating a real stereo pair. The disparity in the matching points (as it is depicted in the “Sparse disparity” column in Figure 10.3) contains coherent depth values. This occurs when the rectification error is very small and there are not false matches.

Having a dual-camera bronchoscope certainly eliminates many of the problems encountered during this experiment. The next step of our analysis, presented in the next section, provides a simulation of a stereo bronchoscopy in virtual environment.

10.2 3D reconstruction in virtual reality

The previous experiment has provided depth clues from images taken by a monocular bronchoscopy. Having explored the anatomy of the bronchial tree, the next experiment will be conducted in virtual environment using a synthetic 3D model allowing for stereoscopic viewing. The usage of virtual reality techniques in clinical applications is getting more wide-spread because of the availability of more detailed simulation models. Virtual simulators offer many advantages to the medical staff: they are especially valuable for training purposes, for pre-operative planning and evaluation of surgical skills [98], [99]. Virtual reality can recreate the conditions of experimental research that would be difficult to propose in the real world. In particular, in this experiment, virtual reality is applied to recreate the typical environment of a bronchoscopy. To this aim, we have preliminarily built a geometric model of a significative segment of the tracheobronchial tree. The synthetic model has been realized with the open source software Blender [100] using real bronchoscopic images as a reference (Figure 10.4). Special care has been taken to

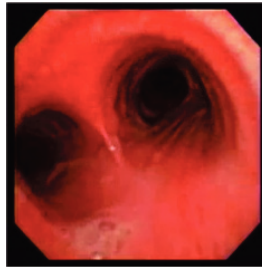


Figure 10.4: A typical real bronchoscopic image.

replicate in the virtual model two of the main geometric parameters of the human pulmonary cavities: the branching degree (i.e. the rate of bifurcation of the air channels as one goes down the respiratory tree) and the decreasing rate of the tube sections after each bifurcation. To enrich with a more realistic value our model, we have reproduced the typical pattern of the respiratory system through the use of some simple procedural textures provided by the standard Blender version. Lighting conditions have been simulated using directional spotlights properly controlled so that light fades down the

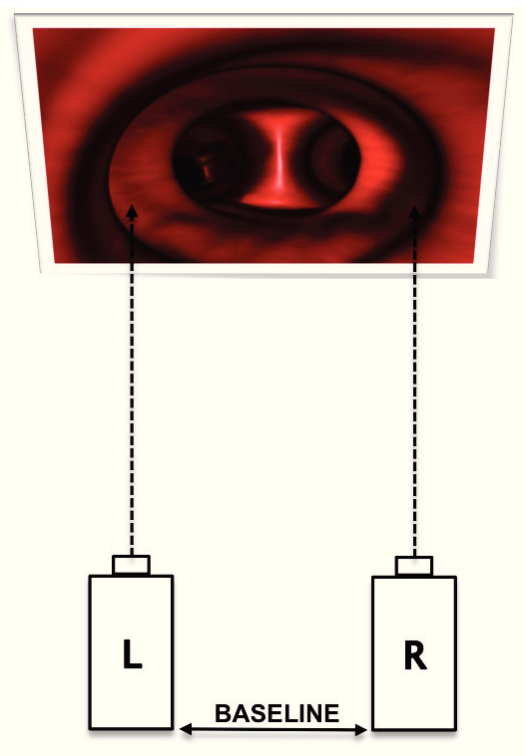


Figure 10.5: Scheme of the canonical stereo system used in virtual reality.

pulmonary branches approximately as the it would in a real bronchoscopy. To perform a stereo reconstruction step in this model, we have placed a couple of aligned cameras with parallel optical axes. The cameras lie in the same vertical and depth axes. They share the same focal length and differ only in the horizontal baseline between them (Figure 10.5). Taking availability of the parameters of the cameras, it is not required to conduct a preliminary calibration step because the images are already rectified. Figure 10.6 shows some pairs of stereo images extracted from the model.

We believe that the ideal conditions provided by virtual reality allow to obtain reliable results, and then use this experiment as a basis for a hypothetical future real application.

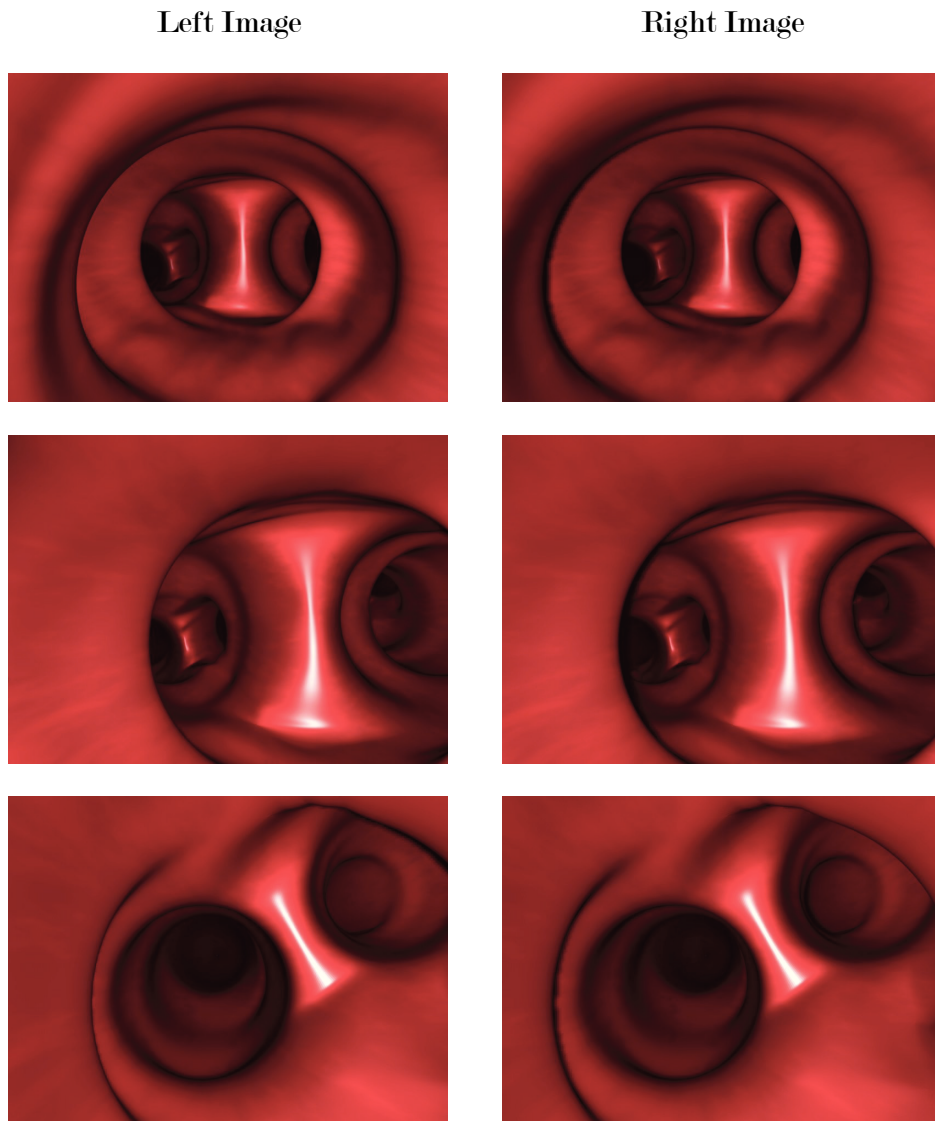


Figure 10.6: Examples of stereo images extracted from the proposed virtual model.

10.2.1 Depth clues extraction

The next step in our analysis involves a stereo recording of a route inside the virtual model. Each rendering is carried out from stereo cameras looking at the scene from two different points of view. The goal is to construct a depth map of the scene from a standard stereo pair acquired by two cameras. In

order to reconstruct the scene we look at the disparity values, i.e., the differences of coordinates of homologous points lying in each image captured by two cameras. As mentioned in Section 8.1, depth is inversely proportional to the disparity; if we represent it such as a gray tone image, brightest pixels correspond to the high values of disparity and consequently to the regions near the cameras. Similarly, the darker pixels represent the deepest ones. By setting focal length and baseline, the depth of a point depends only on the disparity. For the calculation of disparity map we have used the tool in [101]. There are two main motivations for this choice: this method is inspired by algorithms according to the Middlebury stereo evaluation dataset [102]. It also provides the result to support our experimental concepts.

The main problem is to establish which point in right image is the exact projection of the same point in the left image, otherwise known as the correspondences problem. The matching function used in our experiments is inspired by Klaus et al. [103] and consists of a weighted combination of two outcomes: the sum of absolute intensity differences (SAD) and a measure based on gradient determining the disparity by formulating a differential equation which correlates disparity with brightness variations. This match function is defined as:

$$C(x, y, d) = (1 - \omega) * C_{SAD}(x, y, d) + \omega * C_{GRAD}(x, y, d) \quad (10.1)$$

where

$$C_{SAD}(x, y, d) = \sum_{(i,j) \in N(x,y)} |I_1(i, j) - I_2(i + d, j)| \quad (10.2)$$

and

$$C_{GRAD}(x, y, d) = \sum_{(i,j) \in N_x(x,y)} |\nabla_x I_1(i, j) - \nabla_x I_2(i + d, j)| + \sum_{(i,j) \in N_y(x,y)} |\nabla_y I_1(i, j) - \nabla_y I_2(i + d, j)| \quad (10.3)$$

$N(x, y)$ is the correlation window at point (x, y) , ∇_x and ∇_y are the gradients along the horizontal and vertical directions. $N_x(x, y)$ is a correlation window without the rightmost column, $N_y(x, y)$ a correlation window without the

lowest row. The optimal value of disparity d is one that minimizes the match function C . The probability of a wrong match decreases in proportion with the size of the correlation window. A correlation window of size 3×3 pixels is optimal for the reliability of our results. Further parameters needed to estimate disparity assume the default values, as in [101].

Using virtual reality we can easily obtain the field of depth on the rendered images. Then, a qualitative assessment was carried out by comparing our maps with the ground truths depicting real disparities relative to the reference image (left image). Figure 10.7 shows some examples of depth maps calculated with the proposed method. Ground truth images are also reported. As shown in Figure 10.7, depth maps provide an adequate description of the depth of the scene. However, some problems may arise when images have saturated and/or textureless regions. In these circumstances, the amount of correct matches decreases and the resulting disparity map contains inconsistent values.

10.2.2 Augmented Reality

There are different ways to exploit depth information. One of this is Augmented Reality. The final step in our experiments provides for the integration of depth information in the original representation of the scene. All must be optimally developed, in a way that user has the perception of a single scene. To emphasize the effect we take into account depth information in order to meaningfully overlay colors within the image, as proposed in [104], [105]. In detail, red color is associated with the pixels with the maximum value of disparity, corresponding to areas of the scene near the cameras; likewise, blue color is associated with the deeper areas. Intermediate disparity values take gradation colors between red and blue. The colors in red-blue range have a strong impact in the operator than other colors because they are conventionally associated to the situations of danger, warning and safety respectively. In the context of the bronchoscopic images, this representation allows many advantages, including to figure out where objects are lying, so they can be easily avoided. Figure 10.8 shows depth maps integrated in the original rep-

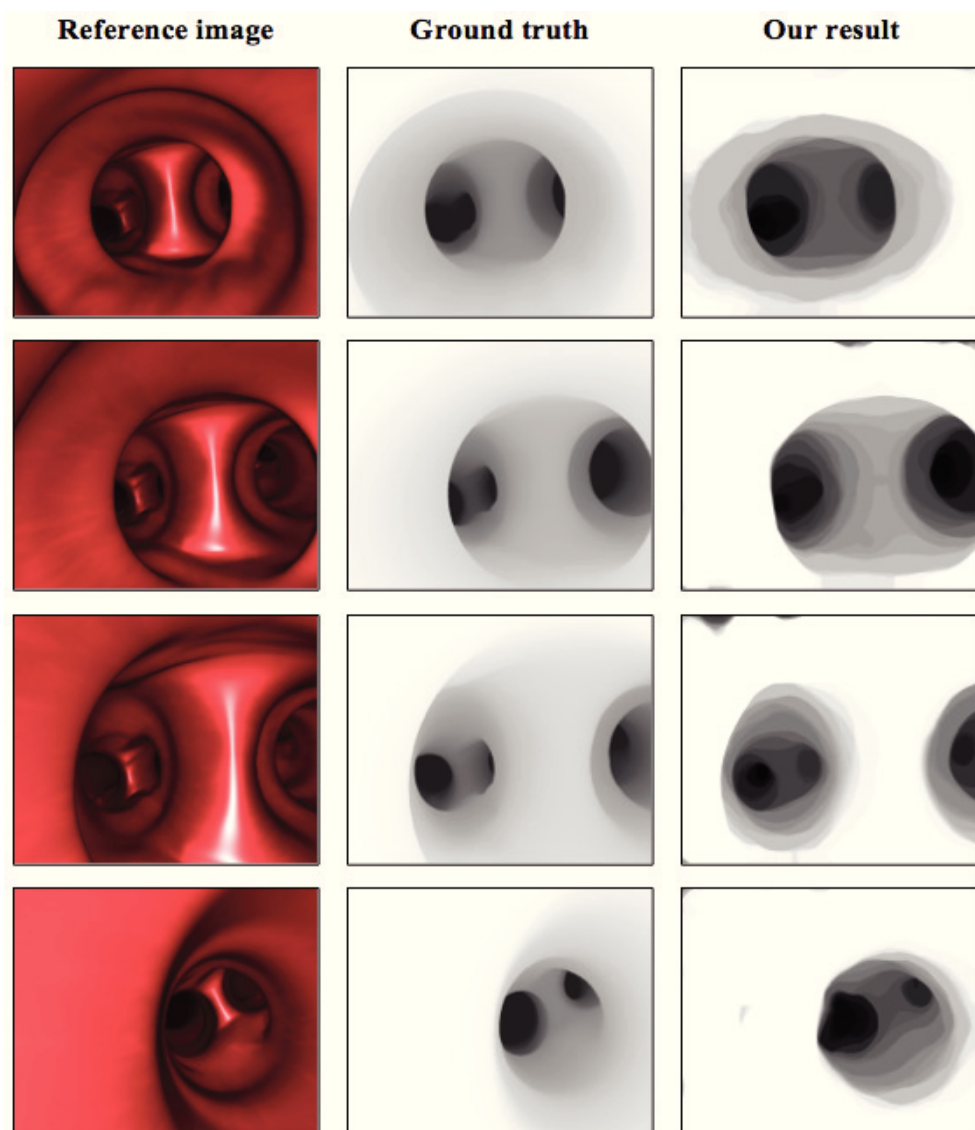


Figure 10.7: Examples of depth maps estimated with the proposed method. The second column shows the ground truth data for the reference images.

representations. We have conducted some tests to verify the percentual of color information to be overlaid in the images. Final results contain the colors that best support our visual investigation and give a greater sense of depth.

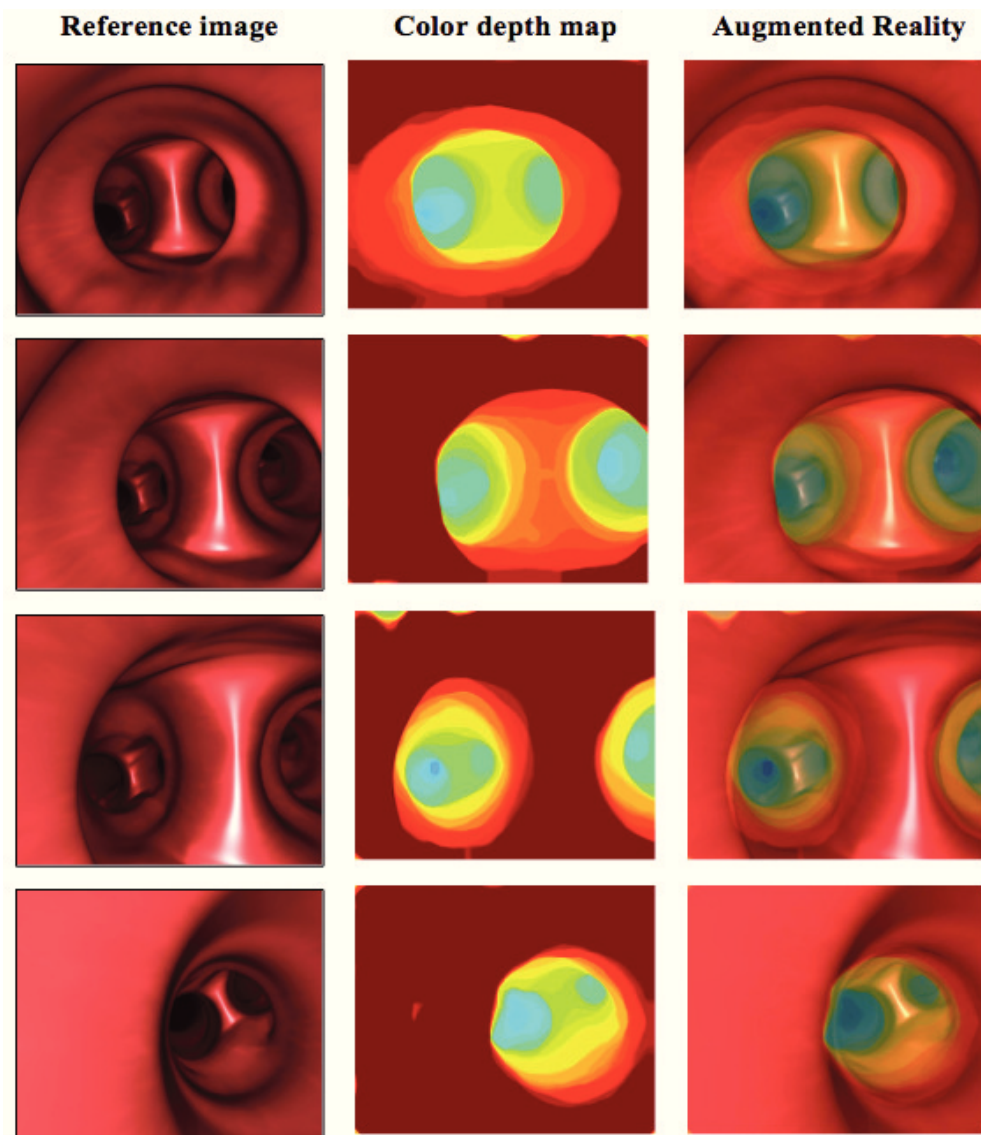


Figure 10.8: Color depth maps integrated in the reference images.

10.2.3 Discussion

Our experiments show that the additional information provided by depth maps leads to a better perception of the distances in the scene. This should in turn likely provide a greater precision in the movements of the bronchoscope, minimizing the number of accidental collisions with the bronchial wall during probe navigation. This last feature provides two main benefits: the

patient undergoes a less discomfort during the examination. Furthermore, the final video contains only meaningful frames to make a good diagnosis. Although at the present time precise data about the effectiveness of the proposed set-up in reducing unwanted collision are unavailable, we believe that the present study supports the application of stereoscopic vision in bronchoscopic applications.

The colored depth map overlaid on the original representation is only one of the potential Augmented Reality visualizations. With this analysis we have to experience with depth map and Augmented Reality visualization based on color in endoscopic context. However, informations that can be integrated on real bronchoscopic images are several. Hence, Augmented Reality in bronchoscopic environment can actually provide an useful instruments to overcome surgeon's perceptual skills.

These tasks may be proposed using only information provided by depth-map images. Additional tools can be developed combining depth informations into a tridimensional mesh, using dense surface stereo reconstruction techniques. In this context, a potential application involves the shape reconstruction, in a post-operative scenario, of route taken by the physician during the examination. In this way the expert can analyze which regions have been explored. This reconstruction can also be used for educational purposes to develop training-oriented systems for the simulation of bronchoscopic examinations. Depth maps in Figure 10.7 provides a detailed description about the depth of the scene. The high number of correct matches is due to the ideal conditions provided by the virtual environment. In the real case, the situation is most likely more challenging as the bronchoscopic images may present a more articulated or smoother texture, which may make harder to solve the correspondence problem. Defocus regions may also be present due to sudden movements by the operator during the navigation. The situation is worsened by the presence of saturated regions of color due to the led light of the surgical probe. In order to obtain the same results proposed in virtual environment, it is necessary to use appropriate denoising Image Processing algorithms. Stereo reconstruction in this kind of images is a difficult issue. The complications are due to the nature of the images that often include ra-

dial distortion. In addition, the matching function used for extract disparity often rely on the use of epipolar rectified images. To overcome these problems, an accurate calibration step is needed in order to obtain information on the perspective view of the scene and to bring images in a standard stereo form.

The problems listed above can be addressed using appropriate Image Processing techniques and the proposed approach can therefore successfully be applied to support endoscopic navigation and intervention. A first step in this direction is to develop a real prototype of a flexible bronchoscope whose tip is equipped with two aligned miniature cameras. This activity has been carried out at the labs of the “School of Engineering and Technology, University of Hertfordshire”. In the next chapter, we give a preview of what has already been done about the flexible stereo bronchoscope.

10.3 Stereoscopic bronchoscope prototype

Modern technologies provide flexible endoscopes that include plenty of accessories and utilities. Currently, there are no companies that offer flexible stereo endoscopes while this solution seems to be promising and it will certainly be soon on the market. A typical problem by using flexible endoscopes is that the operator loses track of the route covered during the navigation. In this regard, some medical tracking systems have been proposed. These systems calculate the position and the orientation of surgical instruments using optical or magnetic sensors [106]. Other approaches attempt to reconstruct the path followed by the endoscope-tip position. However, this does not provide information on tip position at run time. The use of an Augmented Reality interface can revolutionize this field by adding information to the scene to alert the physician if a route has already been covered or to keep track of the depth at which the bronchoscope is currently located.

In this chapter we introduce a prototype of a flexible stereo bronchoscope and discuss our early experience using it to provide depth information that can help the deployment of computer-aided navigation systems. In order to make the proposed system more realistic, we perform our experiments in the bronchial tree of simulation dummies available at the University of Hertfordshire labs. First, we design a complete calibration scheme to estimate both geometric and photometric parameters including spatial and rotation angle of the cameras of the stereo system. Then, we extract depth maps in order to repropose the same Augmented Reality visualization adopted during the simulation in virtual environment (see Section 10.2).

10.3.1 Hardware

The limiting factor in building a bronchoscope is the diameter. The human bronchial tube has a diameter of about 2 cm in its initial section and this gradually decreases as one goes forward in the respiratory tree. To satisfy this requirement we make use of two cameras, each with a diameter of about 5 mm. Each camera also contains four lighting leds whose intensity can be appropriately increased or decreased. The characteristics of the optical

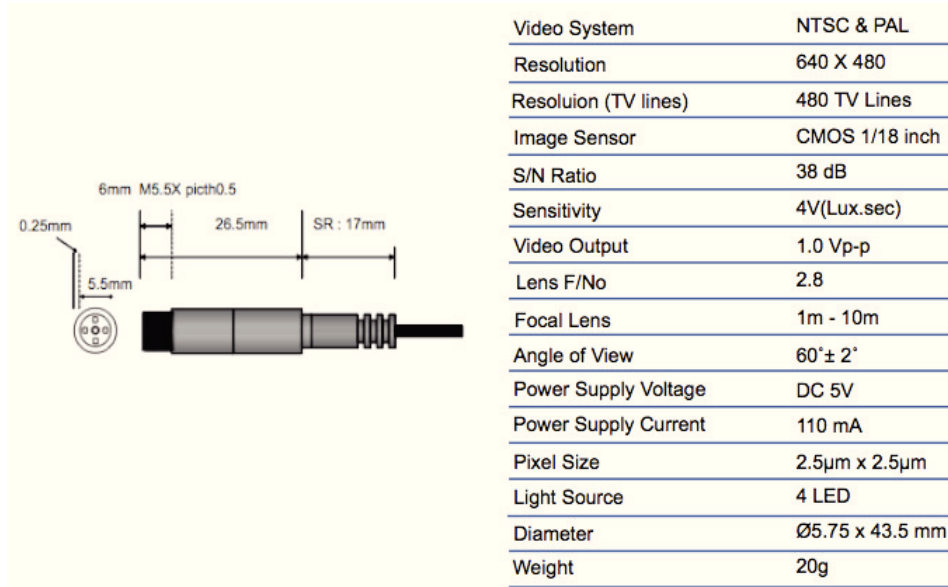


Figure 10.9: Hardware information of the prototype of stereo bronchoscope.

system is reported in Figure 10.9.

The cameras have analog signal and the shooting scene can be displayed in a device equipped with an analog video input. The management with appropriate software is achieved using an analog/digital converter. For the accurate navigation in a tubular surface such as the bronchial tree, we rely on the use of a 1cm-diameter probe whose tip can be articulated in two directions up to 180°. We bought this probe as a fiber-optic system commonly used for the inspection of not easily accessible tubular structures. The ideal would be to have a real bronchoscope, but also having the economic resources, it is yet not possible to find a flexible stereo bronchoscope stereo in the market. For our experiments is enough to have something close to the real case and our probe is readily customizable for the case under examination. Initially, the cameras have been placed on the probe's tip separated by a baseline of 8 mm with parallel optical axes. Further experiments have been made to verify the amount of stereo effect provided for different baselines. The produced stereo effect has been tested by means of stereoscopic displays and 3D glasses [107]. We have achieved the best results when the cameras are perfectly adjacent. This configuration leads to an overall diameter of about 1.5 cm, which is

adequately small for a bronchoscopic examination. Figure 10.10 shows a scheme of how the cameras are cabled for management through computer. The overall stereoscopic system is reported in Figure 10.11.

10.3.2 Software

Calibration of the stereo system

Calibration is a process performed after the capture of the images from the cameras that compose the stereo system. The goal is to accurately measure the intrinsic and extrinsic parameters of the stereo model. With these parameters it is possible to infer information about the coordinates of the points in the real space. There are different techniques to perform the calibration [108]. Typically, this procedure is carried out by using a geometric rig with known geometry, such as a checkerboard pattern. We choose the calibration toolbox developed by Jean-Yves Bouguet and available at [109]. The ease of use and the presence of wide documentation prompted us to use this tool. The toolbox is also written in Matlab, providing broad compatibility with our software. In Bouguet's calibration technique, several images of checkerboards in various positions are captured by each camera simultaneously. Through various perspective views of the checkerboard, the algorithm estimates the position, orientation and internal parameters of each camera. In our experiments we use the checkerboard shown in Figure 10.12(a). We know the number of squares along both horizontal and vertical directions and the size of each square inside the checkerboard pattern ($7mm \times 7mm$). The checkerboard is fixed on a rigid surface not easily deformable. Twenty images are captured for each camera with the checkerboard covering a comprehensive set of positions, rotations and inclinations with respect to the camera that remains fixed (Figure 10.12(b)). The extreme corners of the checkerboard in each captured image are manually located with four mouse clicks by the user, and the toolbox finds the locations of the corners of the internal squares of the checkerboard (Figure 10.13).

Once the corners have been extracted for both sets of left and right images, the algorithm calibrates each camera and then as a stereo pair. Details on

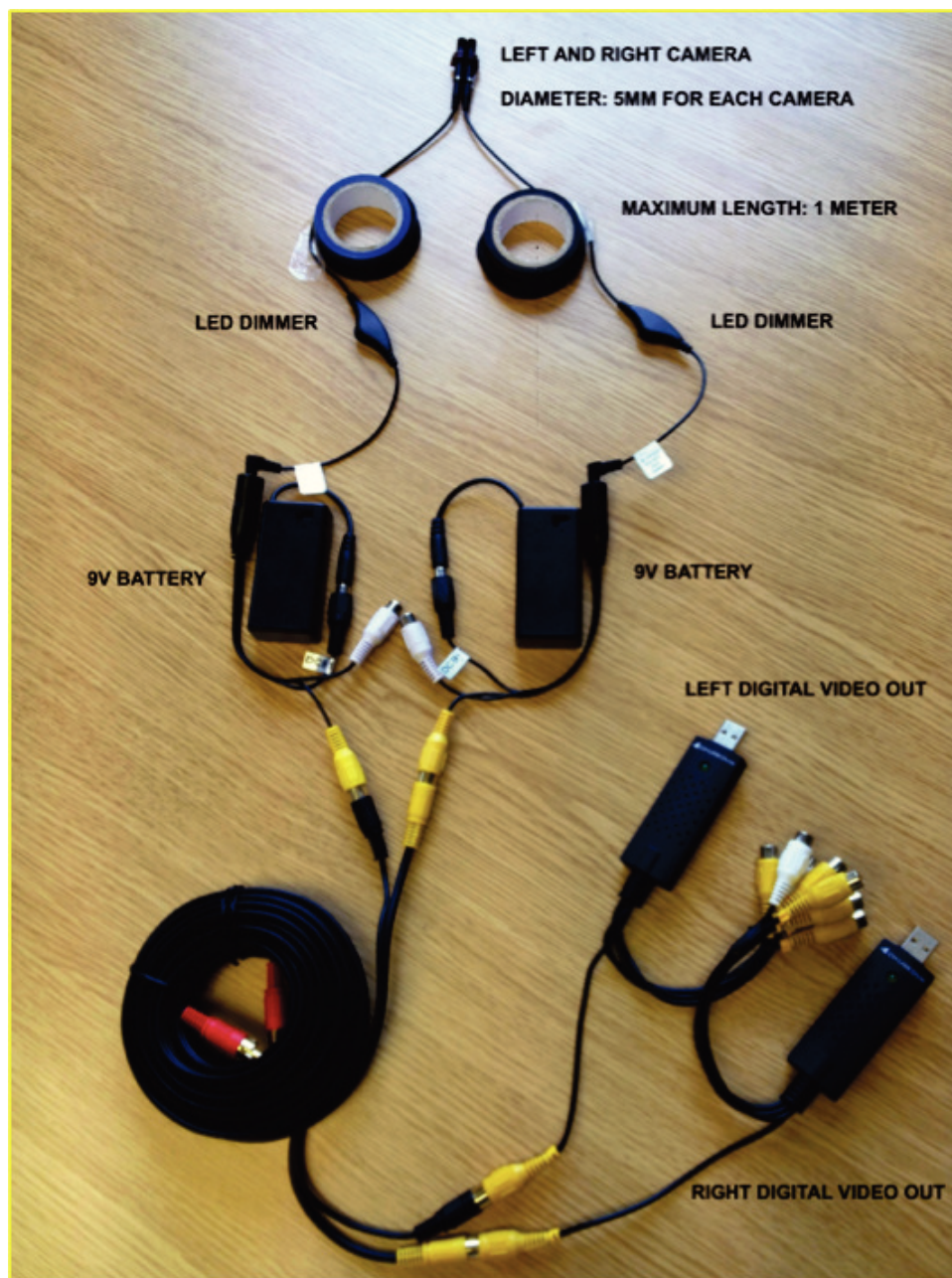


Figure 10.10: Hardware configuration of the prototype of stereo bronchoscope. Two analogic cameras are handled via computer using analog/digital converters. Each camera contains four lighting leds whose intensity can be adjusted with an appropriate dimmer. The whole system is powered by two 9-volt batteries.



Figure 10.11: The prototype of flexible stereo bronchoscope. The cameras are placed at the end of a probe whose tip can be articulated by the user in two directions up to 180° . The best configuration has been obtained with the cameras perfectly adjacent with parallel optical axes.

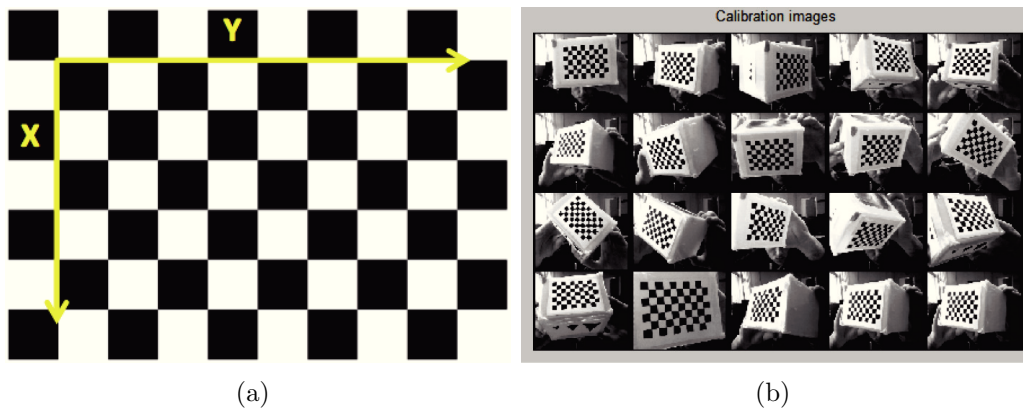


Figure 10.12: (a) Checkerboard pattern used for the calibration of the stereo system. It contains five squares along the vertical axis and eight along the horizontal axis. The size of each square is $7mm \times 7mm$. (b) In order to perform a reliable calibration step, twenty images of the checkerboard placed in different directions and rotation are captured for each camera.

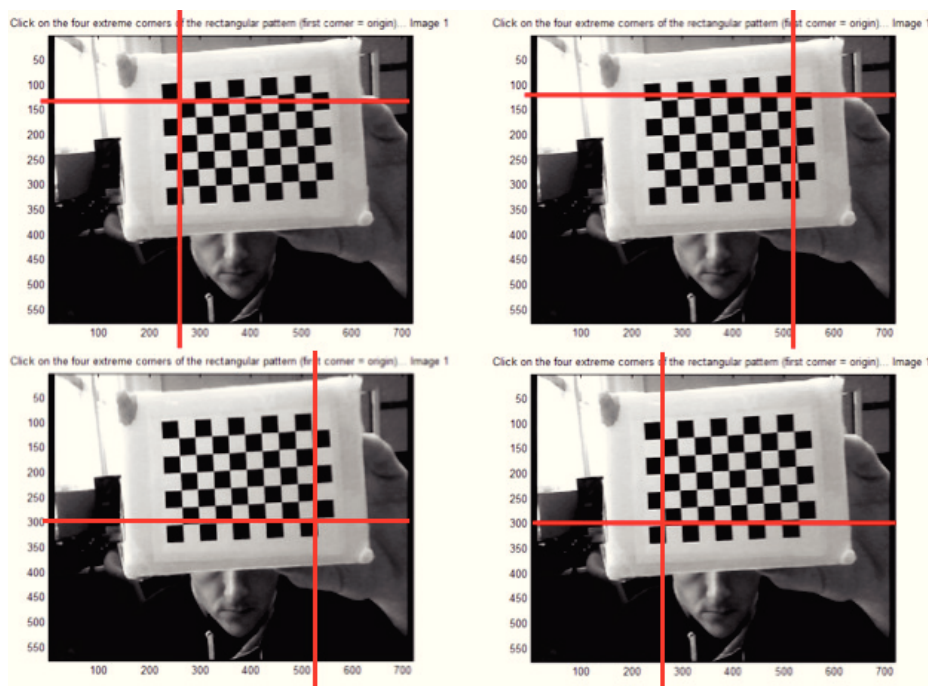


Figure 10.13: Selection of the four angles of the checkerboard. The system estimates the number of squares inside the pattern and tries to extract all the internal corners.

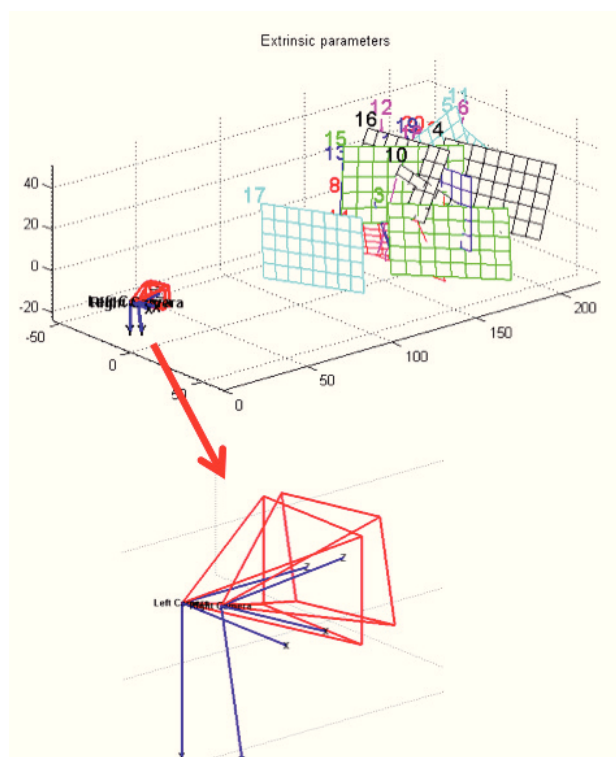


Figure 10.14: Extrinsic parameters of the stereo system. They allow to obtain the spatial configuration of the two cameras respect to the world's reference system.

the mathematical resolution of this procedure are reported in [110]. Each camera produces a set of intrinsic parameters, including the focal length in pixels, the coordinates of the principal point, the possible deformation of the pixels with respect to the ideal square shape (a parameter also known as skew) and the parameters to characterize the distortion caused by the lens. The calibration provides one set of extrinsic parameters for the stereo system with the geometry rotation and translation of the right camera with respect to the left camera. Figure 10.14 illustrates a graphical representation of the extrinsic parameters produced by a stereo calibration over the cameras at our disposal.

An important consequence of the calibration is the rectification of images, i.e., a transformation of the original images in order to produce a stereo pair in which the image planes of the cameras are coplanar and the detection of

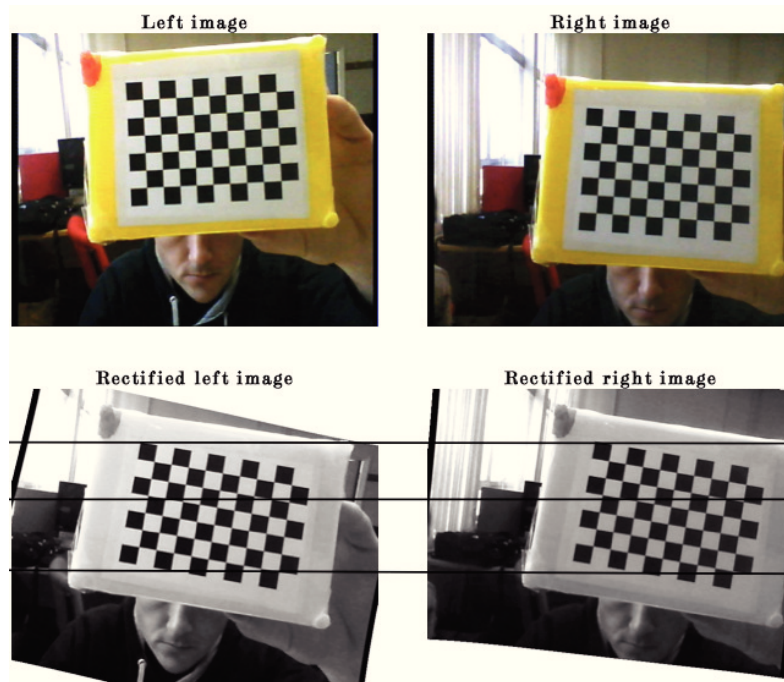


Figure 10.15: Image rectification. During the calibration step, the intrinsic and extrinsic parameters are calculated for each camera. Stereo calibration brings together these parameters allowing for a geometric transformations of the images in a standard stereo form. The black lines superimposed on the rectified images validate the correctness of this procedure.

homologous points is done by examining the same row in two image planes. In figure 10.15 is shown the rectification for a pair of images acquired by our stereo system. In the first row the original images are reported. Rectified left and right images are achieved exploiting the parameters estimate during the calibration. Some horizontal lines are superimposed in the rectified images to verify the accuracy of the procedure.

Acquisition and calculation of disparity map

Figure 10.16 shows a screenshots of the GUI developed for the acquisition of the images from the stereo system. In a stereoscopic system aimed at obtaining three-dimensional information of a dynamic scene, it is necessary that the acquisition of left and right images is simultaneous. Our acquisi-

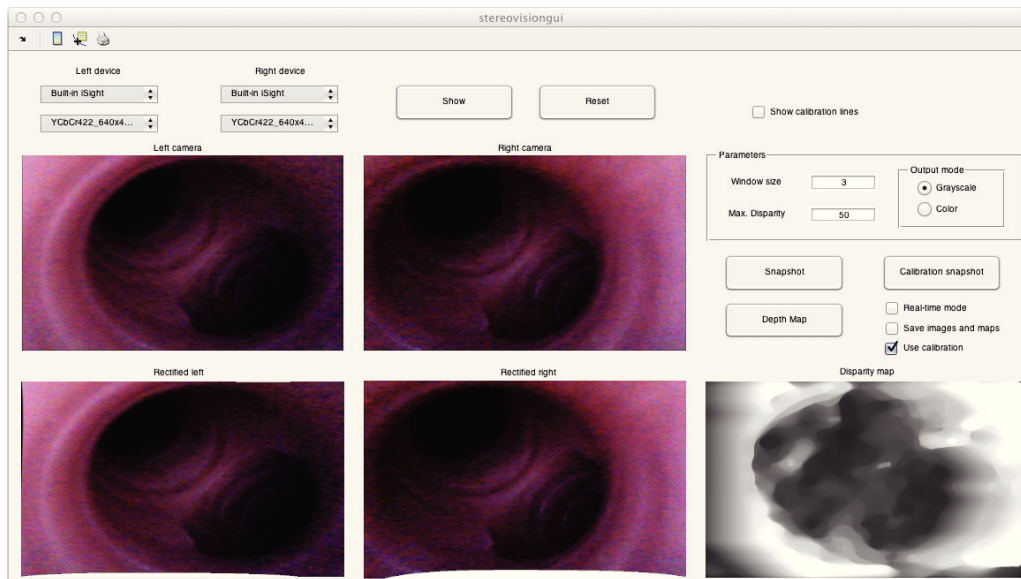


Figure 10.16: GUI of the acquisition software. The user selects the left and right devices and one of the supported resolutions. The system loads the calibration data and rectifies the images captured by the cameras. Disparity map is extracted from rectified images and reported.

tion software detects the devices currently installed on the system. The user selects both left and right cameras and one of the resolution supported by these devices.

The system loads the calibration data and rectifies in real time the images captured by the cameras. Disparity maps are calculated using a dense matching approach with SAD as similarity metric (see Section 8.2). Linear interpolation is used to fill in the “holes” in the disparity maps. These can be computed both for a single pair of rectified images or in real-time mode. However, the calculation in real time of disparity maps only makes sense if small values for the size of the correlation window and maximum disparity are used. If these two parameters are set to high values, the computation may take much longer.

10.3.3 Experimental results

The dataset used in our experiments has been acquired by means of the use of simulation dummies available at the “Clinical Simulation Centre” of the University of Hertfordshire. These are provided by Laerdal company [111]. It produces interactive and realistic simulation dummies for a wide range of medical procedures. The simulator responds to clinical intervention, instructor control, and comprehensive pre-planned scenarios for effective practice of diagnosis and patient care. It also has all the features necessary to the training of hospital staff, with spontaneous breathing, airway control, voice, sounds, and many other clinical features. With realistic anatomy, careful clinical functionality and operation using computer software, it offers numerous educational opportunities for healthcare professionals. Figure 10.17 shows some pictures of the simulation dummy used in our experiments.

The stereoscopic bronchoscopic probe has been previously calibrated by following the procedure described in Section 10.3.2. The imaging device has been set to a resolution of 640×480 pixels working on *RGB* color space.

An important parameter of a stereo system is the baseline separating two camera objectives. A small baseline implies the estimation of small disparities and this may not be enough to ensure a detailed description of the depth of the scene. One could then conclude that it is better to have a high baseline in order to get two significantly different perspectives and therefore high disparities. However, this can be done within certain limits, because with increasing baselines decreases the field of view common to the two cameras and it makes the correspondence problem more difficult. We have conducted experiments with different baseline values. In each of them we noticed that the stereo pairs contain too different images to ensure a reliable reconstruction of the scene. Even when the value of baseline has been reduced to a minimum, making the cameras perfectly adjacent, the situation does not change. Figure 10.18 shows the obtained results. The first column in Figure 10.18 represents the reference image (left image) for the calculation of the correspondence between points. We report in the remaining two columns the disparity maps in the red-blue color range and the Augmented Reality ap-

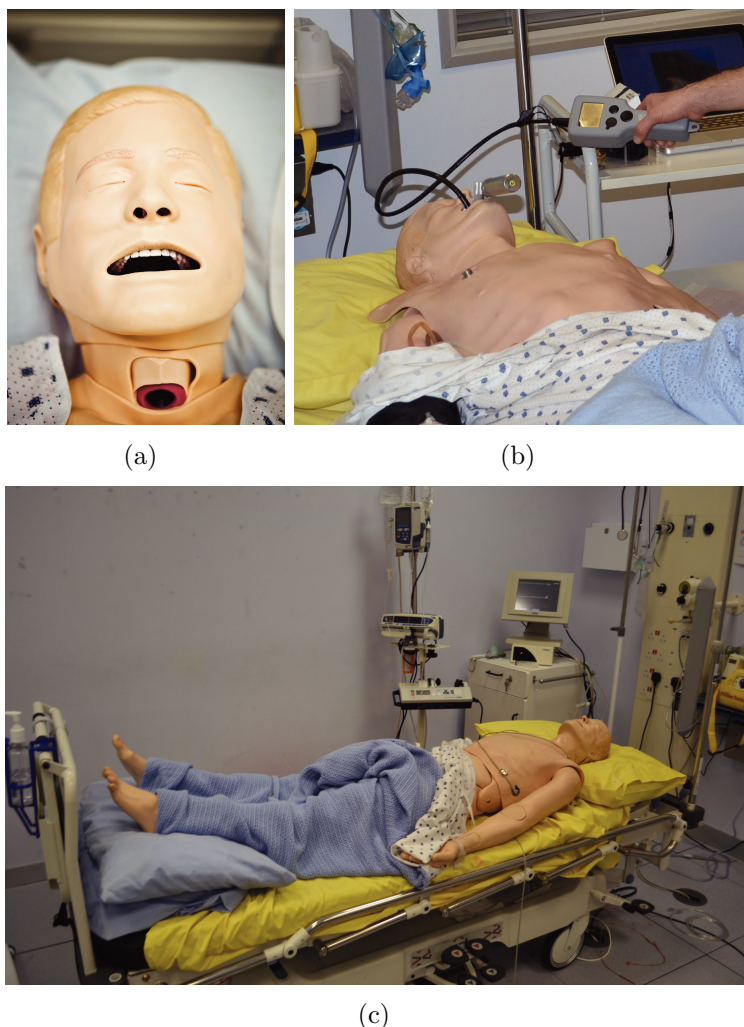


Figure 10.17: The simulation dummy used in the experiments. The insertion of the bronchoscope can occur from the mouth or directly from the trachea. In this case the hole located in the neck of the dummy is exploited. The user observes on a screen the images captured during the bronchoscopic navigation.

plication, obtained as reported in Section 10.2. If the bronchoscope acquires an object located at a depth enough to ensure similar field of views in the two cameras, a reliable disparity map can be achieved. If the bronchoscope is next to an object located at a minimum distance (one or two centimeters), the images that come out are too different to conduct a reliable correspondence analysis and the final disparity map may contain errors.

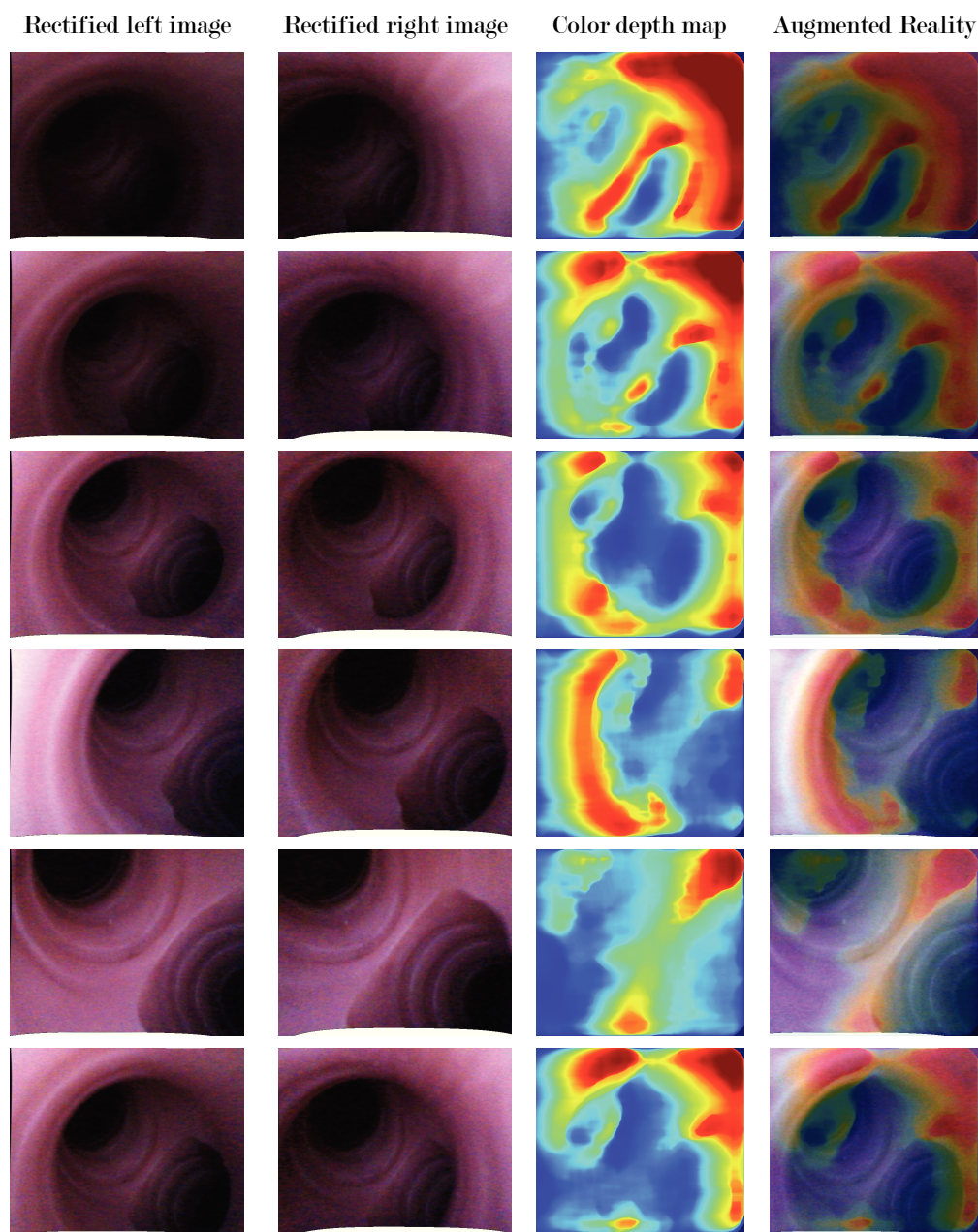


Figure 10.18: Disparity maps obtained using the prototype of flexible stereo bronchoscope. The first two columns show the rectified version of image pairs. Disparity map is reported in the third column using the red-blue range of colors. The last column combines the color disparity map with the reference image (left image).

One of our ongoing activities involves the use of new cameras with a diameter smaller than those currently used and a larger field of view. We believe that, with this new hardware, we can eliminate some of the problems that prevent us from really reliable results.

Chapter 11

Conclusion and future work

The second part of this dissertation has focused on the study of stereo reconstruction algorithms for endoscopic data. Generally, the purpose of these algorithms is to reconstruct the 3D object surface and depth maps. But the main proposal of our analysis is to give useful depth markers to doctors through the use of Augmented Reality. To this aim, we have performed experiments to extract depth maps in three different kinds of endoscopic images. Taking into account the bronchoscopy, i.e., a subset of the endoscopic field, we have conducted a first experiment making use of a bronchoscopic video obtained with a standard monoscopic equipment. In other words, the bronchoscope contains only one camera and the stereo pair required to perform a stereo analysis is constituted by two adjacent images on the video. We have used feature-based matching to perform a calibration step in order to rectify the images in a standard stereo form. This approach strongly depends on the movement of the camera and may properly work only for those image pairs that can “simulate” a real stereo pair. For these images the depth of matching points is estimated and a sparse depth map has been obtained. We have also reported a “semi-dense” representation of the disparity maps obtained making use of a segmentation algorithm in order to detect the main clusters in the image and to fill them with the most representative disparity. In a second experiment we have considered a graphic model of the bronchial tree simulating a stereo bronchoscopy in virtual environment. Unlike the pre-

vious experiment, the depth is here estimated through the use of correlation-based matching. With the availability of ideal stereo images, a subwindow in the left image is paired with its homologous in the right image by using a correlation function. This process repeated for each left subwindow allows to obtain a dense depth map. Once such representation has been obtained, we have proposed an integration of depth information in the original representation of the scene through the use of Augmented Reality. A simple representation that combines the original images and colors in accordance to the depth has been presented to indicate which regions are located near the camera as well as those with greater depth.

In the last experiment we have proposed the same Augmented Reality interface applied in images captured inside the bronchial tree of a medical simulation dummy. To mimic a real bronchoscopy, we have presented a real prototype of flexible stereo bronchoscope. We have reported details about the couple of cameras employed in the bronchoscope and the software used to manage it.

In all the conducted experiments we have represented a depth map as an intensity image that shows in correspondence of each point of the reference image (for example the left image), the value of disparity associated with that point. To facilitate the visualization, the disparity values are mapped by means of a suitable scale factor within the range $[0, 255]$. What we banally call depth map is simply a disparity map. However, it is possible to achieve the actual depth of an object in the scene from its disparity value taking into account the cameras parameters estimated during the calibration step.

Notice that in a simplified use of Augmented Reality, like the one proposed in this dissertation, disparity maps are quite enough to validate our proposal about the use of Augmented Reality in bronchoscopic field. In the general case, the issue is to overlay information to a specific depth in the reference system of the scene. This requirement forces the use of a real depth map in order to preserve the 3D position of physical objects in space and reconstruct the entire three-dimensional structure of the visible scene. Once such a depth representation has been obtained, many Augmented Reality effects may be considered and we suggest their implementation as part of future works. Sev-

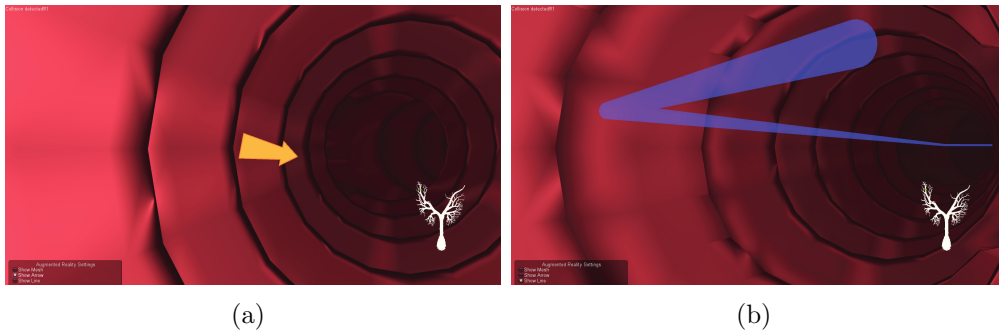


Figure 11.1: Two examples of Augmented Reality effects that can be used in the bronchoscopic context. (a) An arrow marker indicates the proximity to a bronchial wall and the direction to follow to ensure a safe navigation. (b) Another similar effect shows the optimal path to follow during navigation.

eral kinds of markers can be displayed when the bronchoscope is too close to an object, providing additional support to the navigation. Two examples of Augmented Reality effects that can be implemented in the bronchoscopic context are reported in Figure 11.1¹. It is relevant to point out that a correct depth map must be obtained in order to ensure an optimal visualization of these Augmented Reality effects. In this way, a virtual object may be scaled and positioned properly inside the real scene. If the bronchoscope is being moved around the object, the user should observe the virtual object from a different angle.

Although our analysis has involved three different application areas (monoscopic real images, virtual images and images taken inside a medical dummy), it is clear that future works are aimed to improve the results achieved through the use of the bronchoscope prototype. Improving the hardware is one of the activities that we most recommend. The problem of poor lighting in some images can be easily reduced by adding a further lighting led to the bronchoscopic probe. Another improvement concerns the cameras employed in the bronchoscope that produce two totally different points of view when the probe is too close to an object. Having a pair of cameras with smaller diameter and with a larger field of view can help to minimize the baseline between

¹The images in Figure 11.1 are obtained by courtesy of 3D Visualization and Robotics Lab, School EnT, University of Hertfordshire, UK.

the cameras and reduce this problem. With regard to the software implementation, we are in continuous research of newer Computer Vision techniques in order to improve the results achieved by the current methods. However, evaluative experiments proposed in this dissertation show that the proposed system is accurate enough to be used for further studies. Outcomes of this research have indeed relevant implications for the improvement of current endoscopic imaging systems.

Bibliography

- [1] G. Imaging, Expanding the scope of gi (Last accessed: November 2012).
URL <http://www.givenimaging.com>
- [2] Olympus, Medical systems and endoscopy (Last accessed: November 2012).
URL <http://www.olympus-europa.com/endoscopy>
- [3] G. Iddan, A. Glukhovsky, P. Swain, Wireless capsule endoscopy, *Nature* 405 (2000) 725–729.
- [4] A. Culliford, J. Daly, B. Diamond, M. Rubin, P. Green, The value of wireless capsule endoscopy in patients with complicated celiac disease., *Gastrointestinal Endoscopy* 62 (1) (2005) 55–61.
- [5] B. Lewis, P. Swain, Capsule endoscopy in the evaluation of patients with suspected small intestinal bleeding: Results of a pilot study., *Gastrointestinal Endoscopy* 56 (3) (2002) 349–353.
- [6] IntroMedic, Capsule endoscopy has evolved ... mirocam (Last accessed: November 2012).
URL <http://www.intromedic.com>
- [7] S. Bang, J. Y. Park, S. Jeong, Y. H. Kim, H. B. Shim, T. S. Kim, D. H. Lee, S. Y. Song, First clinical trial of the miro capsule endoscope by using a novel transmission technology: electric-field propagation, *Gastrointestinal Endoscopy* 69 (2) (2009) 253 – 259.
- [8] OMOM, Capsule endoscopy (Last accessed: November 2012).
URL <http://english.jinshangroup.com>

- [9] E. Rondonotti, J. Herrerias, M. Pennazio, A. Caunedo, M. Mascarenhas-Saraiva, R. D. Franchis, Complications, limitations, and failures of capsule endoscopy : A review of 733 cases 62 (5) (2005) 712 – 716.
- [10] R. Sidhu, D. S. Sanders, A. J. Morris, M. E. McAlindon, Guidelines on small bowel enteroscopy and capsule endoscopy in adults., *Gut* 57 (1) (2008) 125–36.
- [11] Z. Fireman, E. Mahajna, E. Broide, M. Shapiro, L. Fich, A. Sternberg, Y. Kopelman, E. Scapa, Diagnosing small bowel crohn’s disease with wireless capsule endoscopy., *Gut* 52 (3) (2003) 390–2.
- [12] A. K. H. Chong, A. Taylor, A. Miller, O. Hennessy, W. Connell, P. Desmond, Capsule endoscopy vs. push enteroscopy and enteroclysis in suspected small-bowel crohn’s disease, *Gastrointestinal Endoscopy* 61 (2) (2005) 255–261.
- [13] A. Mata, J. Llach, A. Castells, J. M. Rovira, M. Pellisé, A. Ginès, G. Fernández-Esparrach, M. Andreu, J. M. Bordas, J. M. Piqué, A prospective trial comparing wireless capsule endoscopy and barium contrast series for small-bowel surveillance in hereditary gi polyposis syndromes., *Gastrointestinal Endoscopy* 61 (6) (2005) 721–5.
- [14] W. A. Qureshi, Current and future applications of the capsule camera, *Nature* 3 (2004) 447 – 450.
- [15] F. Arguellas, A. Caunedo, J. Romero, A. Sanchez, M. Tellez, F. Pellicer, F. Arguellas-Martin, J. Herrerias, The value of capsule endoscopy in pediatric patients with a suspicion of chron’s disease, *Endoscopy* 36 (2004) 869 – 873.
- [16] J. Lee, J. Oh, S. K. Shah, X. Yuan, S. J. Tang, Automatic classification of digestive organs in wireless capsule endoscopy videos, in: *Proceedings of the 2007 ACM symposium on Applied computing, SAC '07*, ACM, New York, NY, USA, 2007, pp. 1041–1045.

- [17] M. Mackiewicz, J. Berens, M. Fisher, Wireless capsule endoscopy color video segmentation, *Medical Imaging, IEEE Transactions on* 27 (12) (2008) 1769–1781.
- [18] J. Berens, M. Mackiewicz, G. Bell, Stomach, intestine, and colon tissue discrimination for wireless capsule endoscopy images, in: *SPIE Medical Imaging 2005: Image Processing*, Vol. 5747, 2005, pp. 283–290.
- [19] M. Mackiewicz, J. Berens, M. Fisher, G. Bell, Using colour distributions to discriminate tissues in wireless capsule endoscopy images, *Medical Imaging Understanding and Analyses Conference (MIUA '05)*.
- [20] M. Mackiewicz, J. Berens, M. Fisher, D. Bell, Colour and texture based gastrointestinal tissue discrimination, in: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Vol. 2, 2006, p. II.
- [21] M. Mackiewicz, J. Berens, M. Fisher, Wireless capsule endoscopy video segmentation using support vector machine and hidden markov models, *Medical Imaging Understanding and Analyses Conference (MIUA '06)*.
- [22] J. Berens, M. Mackiewicz, G. Bell, C. Jamieson, Can we detect when a wireless capsule endoscope leaves the stomach using computational colour techniques? a pilot study, *Endoscopy* 36 (1).
- [23] M. Mackiewicz, J. Berens, M. Fisher, G. Bell, C. Jamieson, Computational colour techniques can speed up the viewing of wireless capsule endoscopy images as well as determine gastric and intestinal transit times, *Endoscopy* 54 (2).
- [24] M. Coimbra, P. Campos, J. Cunha, Extracting clinical information from endoscopic capsule exams using mpeg-7 visual descriptors, *IEE Seminar Digests 2005 (11099)* (2005) 105–110.
- [25] M. T. Coimbra, J. P. da Silva Cunha, Mpeg-7 visual descriptors - contributions for automated feature extraction in capsule endoscopy., *IEEE Trans. Circuits Syst. Video Techn.* 16 (5) (2006) 628–637.

- [26] J. P. da Silva Cunha, M. T. Coimbra, P. Campos, J. M. Soares, Automated topographic segmentation and transit time estimation in endoscopic capsule exams., *IEEE Trans. Med. Imaging* 27 (1) (2008) 19–27.
- [27] M. Coimbra, P. Campos, J. Cunha, Topographic segmentation and transit time estimation for endoscopic capsule exams, in: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Vol. 2, 2006, p. II.
- [28] S. F. Chang, T. Sikora, A. Puri, Overview of the MPEG-7 standard, *IEEE Trans. Circuits and Systems for Video Technology* 11 (6) (2001) 688–695.
- [29] M. T. Coimbra, J. Kustra, P. Campos, J. P. da Silva Cunha, Combining color with spatial and temporal position of the endoscopic capsule for improved topographic classification and segmentation, in: *SAMT (Posters and Demos)*, Vol. 233 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2006.
- [30] N. Marques, E. Dias, J. P. S. Cunha, M. Coimbra, Compressed domain topographic classification for capsule endoscopy., *Conf Proc IEEE Eng Med Biol Soc 2011* (2011) 6631–4.
- [31] E. M. Quigley, Gastric and small intestinal motility in health and disease., *Gastroenterol Clin North Am* 25 (1) (1996) 113–45.
- [32] P. Spyridonos, F. Vilariño, J. Vitrià, F. Azpiroz, P. Radeva, Anisotropic feature extraction from endoluminal images for detection of intestinal contractions, in: *Proceedings of the 9th international conference on Medical Image Computing and Computer-Assisted Intervention - Volume Part II, MICCAI'06*, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 161–168.
- [33] F. Vilarino, P. Spyridonos, O. Pujol, J. Vitria, P. Radeva, Automatic detection of intestinal juices in wireless capsule video endoscopy, in:

- Proceedings of the 18th International Conference on Pattern Recognition - Volume 04, ICPR '06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 719–722.
- [34] C. Signorelli, F. Villa, E. Rondonotti, C. Abbiati, G. Beccari, R. de Franchis, Sensitivity and specificity of the suspected blood identification system in video capsule enteroscopy 37 (12) (2005) 1170 – 1173.
- [35] S. Hwang, J. Oh, J. Cox, S. J. Tang, H. F. Tibbals, Blood detection in wireless capsule endoscopy using expectation maximization clustering, in: J. M. Reinhardt, J. P. W. Pluim (Eds.), Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 6144, 2006, pp. 577–587.
- [36] P. Y. Lau, P. Correia, Detection of bleeding patterns in wce video using multiple features, in: Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, 2007, pp. 5601–5604.
- [37] B. Li, M. Q.-H. Meng, Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments, Computers in Biology and Medicine 39 (2) (2009) 141–147.
- [38] M. Boulougoura, V. Wadge, V. S. Kodogiannis, H. S. Chowdrey, Intelligent systems for computer-assisted clinical endoscopic image analysis.
- [39] B. Li, M. Q.-H. Meng, Texture analysis for ulcer detection in capsule endoscopy images, Image and Vision Computing 27 (9) (2009) 1336–1342.
- [40] S. Bejakovic, R. Kumar, T. Dassopoulos, G. Mullin, G. Hager, Analysis of crohn's disease lesions in capsule endoscopy images, in: Robotics and Automation, 2009. ICRA '09. IEEE International Conference on, 2009, pp. 2793–2798.

- [41] B. Li, M. Q.-H. Meng, J. Y. Lau, Computer-aided small bowel tumor detection for capsule endoscopy, *Artificial Intelligence in Medicine* 52 (1) (2011) 11–16.
- [42] P. Grünwald, P. M. B. Vitányi, Shannon information and kolmogorov complexity, CoRR cs.IT/0410002.
- [43] K. J. Balakrishnan, N. A. Touba, Relating entropy theory to test data compression, in: *Proceedings of the European Test Symposium, Ninth IEEE, ETS '04, IEEE Computer Society, Washington, DC, USA, 2004*, pp. 94–99.
- [44] R. Solomonoff, *Information and Control*.
- [45] A. Kolmogorov, *Problems of Information and Transmission*.
- [46] G. Chaitin, *Journal of the Association for Computing Machinery*.
- [47] M. Li, P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer, 1997.
- [48] C. Bennett, P. Gacs, P. Vitanyi, W. H. Zurek, Information distance, *IEEE Trans. Information Theory* 44 (1998) 1407–1423.
- [49] R. Cilibrasi, P. Vitanyi, Clustering by compression, *IEEE Trans. Information Theory* 51 (2005) 1523–1545.
- [50] M. Li, X. Chen, P. Vitanyi, The similarity metric, *IEEE Trans. Information Theory* (2004) 3250–3264.
- [51] T. G. Dietterich, Ensemble methods in machine learning, in: *Proceedings of the First International Workshop on Multiple Classifier Systems, MCS '00, Springer-Verlag, London, UK, UK, 2000*, pp. 1–15.
- [52] E. Bauer, R. Kohavi, An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, *Machine Learning* 36 (1999) 105–139.

- [53] T. G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, in: *Machine Learning*, 1998, pp. 139–157.
- [54] D. Opitz, R. Maclin, Popular ensemble methods: An empirical study, *Journal of Artificial Intelligence Research* 11 (1999) 169–198.
- [55] M. J. Kearns, U. V. Vazirani, *An introduction to computational learning theory*, MIT Press, Cambridge, MA, USA, 1994.
- [56] L. G. Valiant, A theory of the learnable, *Commun. ACM* 27 (11) (1984) 1134–1142.
- [57] R. E. Schapire, The strength of weak learnability, *Machine Learning* 5 (2) (1990) 197–227.
- [58] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: *Proceedings of the Second European Conference on Computational Learning Theory*, Springer-Verlag, London, UK, 1995, pp. 23–37.
- [59] H. Drucker, C. Cortes, Boosting decision trees, in: *Advances in Neural Information Processing Systems (NIPS 1995)*, 1995, pp. 479–485.
- [60] Y. Freund, R. E. Schapire, Discussion of the paper “arcing classifiers” by leo breiman, *Annals of Statistics* 26 (1998) 824–832.
- [61] P. Viola, M. Jones, Robust real-time object detection, *International Journal of Computer Vision* 57 (2) (2002) 137–154.
- [62] C. Papageorgiou, M. Oren, T. Poggio, A general framework for object detection, in: *Computer Vision, 1998. Sixth International Conference on*, 1998, pp. 555 – 562.
- [63] F. C. Crow, Summed-area tables for texture mapping, in: *Proceedings of the 11th annual conference on Computer graphics and interactive techniques, SIGGRAPH '84*, ACM, New York, NY, USA, 1984, pp. 207–212.

- [64] G. Gallo, E. Granata, A. Torrìsi, Information theory based wce video summarization, in: Pattern Recognition (ICPR), 2010 20th International Conference on, 2010, pp. 4198–4201.
- [65] G. Gallo, A. Torrìsi, Boosted wireless capsule endoscopy frames classification, in: Proc. of Third International Conferences on Pervasive Patterns and Applications, PATTERNS'11, pp. 25–30.
- [66] G. Gallo, A. Torrìsi, Lumen detection in endoscopic images: a boosting classification approach, International Journal On Advances in Intelligent Systems 5 (2012) 127–134.
- [67] G. Gallo, A. Torrìsi, Random forests based wce frames classification, in: Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on, 2012, pp. 1–6.
- [68] T. Leung, J. Malik, Representing and recognizing the visual appearance of materials using three-dimensional textons, International Journal of Computer Vision 43 (1) (2001) 29–44.
- [69] M. Varma, A. Zisserman, A statistical approach to texture classification from single images, International Journal of Computer Vision 62 (1-2) (2005) 61–81.
- [70] S. Battiato, G. M. Farinella, G. Gallo, D. Ravì, Exploiting textons distributions on spatial hierarchy for scene classification, J. Image Video Process. 2010 (2010) 7:1–7:13.
- [71] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd Edition, Wiley, New York, 2001.
- [72] G. Gallo, E. Granata, G. Scarpulla, Sudden changes detection in WCE video, International Conference on Image Analysis and Processing 5716 (2009) 701–710.
- [73] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by probability distributions, Bull. Calcutta Math. Soc. 35 (1943) 99–109.

- [74] A. Bardera, M. Feixas, I. Boada, M. Sbert, Compression-based Image Registration, IEEE International Symposium on Information Theory (2006) 436–440.
- [75] A. Kaltchenko, Algorithms for estimating information distance with application to bioinformatics and linguistics, CoRR cs.CC/0404039.
- [76] G. Gallo, E. Granata, G. Scarpulla, Wireless Capsule Endoscopy video segmentation, IEEE International Workshop on Medical Measurements and Applications (2009) 236–340.
- [77] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Computer Vision and Pattern Recognition, CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, Vol. 1, 2001, pp. 511–518.
- [78] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
- [79] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth and Brooks, Monterey, CA, 1984.
- [80] L. Breiman, Bagging predictors, in: Machine Learning, 1996, pp. 123–140.
- [81] L. W. Way, L. Stewart, W. Gantert, K. Liu, C. M. Lee, K. Whang, J. G. Hunter, Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective, Annals of surgery 237 (4) (2003) 460–469.
- [82] J. Leven, D. Burschka, R. Kumar, G. Zhang, S. Blumenkranz, X. Dai, M. Awad, G. D. Hager, M. Marohn, M. Choti, et al., Davinci canvas: A telerobotic surgical system with integrated, robot-assisted, laparoscopic ultrasound capability, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI 2005), Vol. 3749 of Lecture Notes in Computer Science, Springer, 2005, pp. 811–818.
- [83] VisionSense, Stereo vision endoscope (Last accessed: November 2012). URL <http://www.visionsense.com>

- [84] K. K. Badani, A. Bhandari, A. Tewari, M. Menon, Comparison of two-dimensional and three-dimensional suturing: is there a difference in a robotic surgery setting?, *J Endourol* 19 (10) (2005) 1212–5.
- [85] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, *International Journal of Computer Vision* (1-3) (2002) 7–42.
- [86] F. Devernay, F. Mourgues, E. Coste-Mani ere, Towards endoscopic augmented reality for robotically assisted minimally invasive cardiac surgery, in: *Proceedings of the International Workshop on Medical Imaging and Augmented Reality (MIAR 2001)*, IEEE Computer Society, Washington, DC, USA, 2001.
- [87] F. Devernay, 3d reconstruction of the operating field for image overlay in 3d-endoscopic surgery, in: *Proceedings of the IEEE and ACM International Symposium on Augmented Reality (ISAR'01)*, IEEE Computer Society, Washington, DC, USA, 2001, p. 191.
- [88] W. W. Lau, N. A. Ramey, J. J. Corso, N. V. Thakor, G. D. Hager, Stereo-based endoscopic tracking of cardiac surface deformation., in: *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2004)*, Vol. 3217 of *Lecture Notes in Computer Science*, Springer, 2004, pp. 494–501.
- [89] D. Stoyanov, M. Scarzanella, P. Pratt, G.-Z. Yang, Real-time stereo reconstruction in robotically assisted minimally invasive surgery, in: *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2010)*, Vol. 6361 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2010, pp. 275–282.
- [90] A. F. Durrani, G. M. Preminger, Three-dimensional video imaging for endoscopic surgery, *Computers in Biology and Medicine* 25 (2) (1995) 237–247.
- [91] S. Ota, D. Deguchi, T. Kitasaka, K. Mori, Y. Suenaga, Y. Hasegawa, K. Imaizumi, H. Takabatake, M. Mori, H. Natori, Augmented display

- of anatomical names of bronchial branches for bronchoscopy assistance, in: Proceedings of the 4th international workshop on Medical Imaging and Augmented Reality, MIAR '08, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 377–384.
- [92] D. Deguchi, K. Ishitani, T. Kitasaka, K. Mori, Y. Suenaga, H. Takabatake, M. Mori, H. Natori, A method for bronchoscope tracking using position sensor without fiducial markers, in: SPIE Medical Imaging, Vol. 6511, 2007, pp. 65110N–65110N–12.
- [93] A. Torrisi, S. Livatino, G. Gallo, 3d reconstruction and augmented reality in bronchoscopic intervention, in: Eurographics Italian Chapter Conference, 2011, pp. 9–13.
- [94] A. Fusiello, Epipolar rectification (Last accessed: November 2012).
URL <http://www.diegm.uniud.it/fusiello/demo/rect/>
- [95] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision (IJCV)* 60 (2) (2004) 91–110.
- [96] M. A. Fischler, R. C. Bolles, Readings in computer vision: issues, problems, principles, and paradigms, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987, Ch. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, pp. 726–740.
- [97] A. Fusiello, L. Irsara, Quasi-euclidean uncalibrated epipolar rectification, in: International Conference on Pattern Recognition (ICPR), 2008, pp. 1–4.
- [98] O. Van Der Meijden, M. Schijven, The value of haptic feedback in conventional and robot-assisted minimal invasive surgery and virtual reality training: a current review, *Surgical Endoscopy* 23 (2009) 1180–1190.
- [99] P. Heng, P. Fung, K. Sak Leung, H. Qiu Sun, T. Wong, Virtual bronchoscopy, *The International Journal of Virtual Reality*, 4 (4).

- [100] Blender, free open source 3d content creation suite (Last accessed: November 2012).
URL <http://www.blender.org>
- [101] Lankton, Fast 3d stereo vision (Last accessed: November 2012).
URL <http://www.shawnlankton.com/2008/04/stereo-vision-update-with-new-code>
- [102] H. Hirschmüller, Evaluation of cost functions for stereo matching, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.
- [103] A. Klaus, M. Sormann, K. Karner, Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure, in: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, Vol. 3, 2006, pp. 15–18.
- [104] S. Livatino, G. Muscato, D. De Tommaso, M. Macaluso, Augmented reality stereoscopic visualization for intuitive robot teleguide, in: Industrial Electronics (ISIE), 2010 IEEE International Symposium on, 2010, pp. 2828 –2833.
- [105] R. Williams, The non-designer’s design book: design and typographic principles for the visual novice, Peachpit Press, Berkeley, CA, USA, 1994.
- [106] NDI, 3d real-time measurement enabling computer-assisted surgery and therapy (Last accessed: November 2012).
URL <http://www.ndigital.com/medical>
- [107] Nvidia, 3d glasses and displays (Last accessed: November 2012).
URL <http://www.nvidia.com/object/3d-vision-glasses.html>
- [108] E. Trucco, A. Verri, Introductory Techniques for 3-D Computer Vision, Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998.
- [109] J. Y. Bouguet, Camera calibration toolbox for matlab (Last accessed: November 2012).
URL http://www.vision.caltech.edu/bouguetj/calib_doc/index.html

- [110] Z. Zhang, Flexible camera calibration by viewing a plane from unknown orientations, in: *International Journal of Computer Vision (IJCV)*, 1999, pp. 666–673.
- [111] Laerdal, Helping save lives (Last accessed: November 2012).
URL <http://www.laerdal.com/>