Computer-assisted Wireless Capsule Endoscopy video analysis

Michał Mackiewicz

A thesis submitted for the Degree of Doctor of Philosophy

University of East Anglia
School of Computing Sciences

December 3, 2007

©This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that no quotation from the thesis, nor any information derived therefrom, may be published without the author’s prior written consent.
Abstract

Wireless Capsule Endoscopy (WCE) is a relatively recent technology that enables imaging of the entire human Gastrointestinal (GI) tract. It is particularly suited for computer-assisted diagnosis as it records a large number of images from the GI tract, which consequently require a time-consuming visual assessment that can only be carried out by a trained clinician. The duration of this assessment typically varies from one to two hours. Thus, it is clear that in terms of time requirement, the WCE is one of the most costly imaging procedures. Clearly, there is role for automatic analysis using computer vision techniques, in order to aid the analysis of the WCE footage, and thus reduce the time required to reach the diagnosis and reduce the cost of the procedure.

This thesis examines the problems of Wireless Capsule Endoscopy video analysis from an image processing perspective. In particular, we focus on two areas: topographic video segmentation and bleeding detection. The former considers the problem of WCE video segmentation into anatomical organs such as mouth, oesophagus, small intestine and colon. In order to achieve this, we investigated a wide range of computer vision and image processing techniques. We showed that the best performing methods can provide very good segmentation results almost matching a human expert.

The second area that is described in this thesis is bleeding detection. Here, the aim is to automatically find those frames in the WCE video that contain blood. We propose a novel algorithm capable of performing this task. Our two stage method consists of a adaptive histogram based pixel classifier and a region and neighbourhood classi-
In a detailed study, we show that our method outperforms state-of-art commercial software.
Acknowledgements
# Contents

Abstract i  
Acknowledgements iii  
List of figures xi  
List of tables xii  
Publications xiii  
Glossary xv  

1 Introduction 1  
1.1 Why does Wireless Capsule Endoscopy video analysis need computer assistance? 1  
1.2 Main areas of research contained in the thesis 2  
1.3 Thesis organisation 4  
1.3.1 Thesis outline 5  

2 WCE technology and its clinical importance 6  
2.1 Technology 6  
2.1.1 Future development directions 13  
2.2 WCE clinical importance 14  
2.3 Clinical need for better computational tools 19  
2.4 Conclusions 21  

3 Literature Review 22  
3.1 Application of computer vision in WCE 22  
3.1.1 Intestinal fluid detection 23  
3.1.2 Abnormality detection 24  
3.1.3 Capsule speed estimation/retention detection 25  
3.1.4 Detection of intestinal contractions 25
3.1.5 Adaptive viewing speed adjustment .......................... 26
3.1.6 Image quality enhancement ................................. 28
3.2 Application of Computer Vision in conventional endoscopy ... 28
  3.2.1 Colonoscopy image analysis .............................. 29
  3.2.2 Endoscopy image analysis ............................... 31
3.3 Possibilities of image retrieval in a database of endoscopic images 33
3.4 Conclusions ......................................................... 34

4 Topographic video segmentation 37
  4.1 Motivation ......................................................... 37
  4.2 Video regions and their boundaries .......................... 40
  4.3 Related work ..................................................... 41
  4.4 Feature extraction ............................................. 42
    4.4.1 Colour ....................................................... 42
    4.4.2 Texture ..................................................... 43
    4.4.3 Motion ....................................................... 47
    4.4.4 Sub-image region (SubIR) selection ..................... 52
    4.4.5 Compression ............................................... 55
  4.5 Single image classification .................................. 60
    4.5.1 K Nearest Neighbour (kNN) ............................ 61
    4.5.2 Multivariate Gaussian .................................... 62
    4.5.3 Support Vector Classifier ............................... 63
  4.6 Video segmentation .......................................... 69
    4.6.1 A naïve segmentation algorithm based on converging search 70
    4.6.2 Sliding Window ........................................... 70
    4.6.3 Hidden Markov Model ..................................... 72
  4.7 Single image classification experiments ..................... 74
    4.7.1 Results ...................................................... 76
  4.8 Video segmentation experiments ................................ 80
    4.8.1 Results ...................................................... 83
  4.9 Prototype ......................................................... 88
  4.10 Conclusions ...................................................... 89

5 Bleeding detection 91
  5.1 Motivation ....................................................... 92
  5.2 Related work .................................................... 93
  5.3 Method overview .............................................. 94
  5.4 Adaptive histogram based pixel classifier .................. 96
    5.4.1 Colour space choice, Prior histogram learning ........ 96
    5.4.2 Histogram Adaptation .................................... 97
    5.4.3 Pixel classification ....................................... 100
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4.4 Finding optimal adaptation coefficients</td>
<td>100</td>
</tr>
<tr>
<td>5.5 Classification of detected regions</td>
<td>103</td>
</tr>
<tr>
<td>5.5.1 Feature extraction</td>
<td>106</td>
</tr>
<tr>
<td>5.5.2 Region classification experiment</td>
<td>111</td>
</tr>
<tr>
<td>5.5.3 Results</td>
<td>113</td>
</tr>
<tr>
<td>5.5.4 Comparison with Given SBI</td>
<td>116</td>
</tr>
<tr>
<td>5.6 Prototype</td>
<td>124</td>
</tr>
<tr>
<td>5.7 Conclusions</td>
<td>126</td>
</tr>
<tr>
<td>6 Conclusions</td>
<td>129</td>
</tr>
<tr>
<td>6.1 Thesis contributions</td>
<td>129</td>
</tr>
<tr>
<td>6.2 Current state-of-art and future work</td>
<td>132</td>
</tr>
<tr>
<td>A Signed rank test results</td>
<td>137</td>
</tr>
<tr>
<td>A.1 Wilcoxon signed rank test</td>
<td>137</td>
</tr>
<tr>
<td>A.2 Wilcoxon signed rank test results of topographic video segmentation</td>
<td>138</td>
</tr>
<tr>
<td>A.2.1 Naïve (1 pass) versus Naïve (multiple passes)</td>
<td>139</td>
</tr>
<tr>
<td>A.2.2 Naïve versus Sliding Window</td>
<td>143</td>
</tr>
<tr>
<td>A.2.3 Sliding Window versus HMM</td>
<td>147</td>
</tr>
<tr>
<td>A.2.4 Multivariate Gaussian versus SVC</td>
<td>153</td>
</tr>
<tr>
<td>A.2.5 Entire image features versus entire image plus sub-image region features</td>
<td>158</td>
</tr>
<tr>
<td>A.2.6 Different features</td>
<td>162</td>
</tr>
<tr>
<td>References</td>
<td>167</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Given Imaging M2A capsule</td>
<td>8</td>
</tr>
<tr>
<td>2.2</td>
<td>The components of the M2A capsule - 1) Optical dome, 2) Lens holder,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) Lens, 4) Illuminating LEDs, 5) CMOS imager, 6) Battery, 7) ASIC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RF transmitter, 8) Antenna</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>Rapid Reader ver.4 software - A part of the Given Imaging diagnostic system.</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Note, the quad view - four images blended on the edges and displayed at the</td>
<td></td>
</tr>
<tr>
<td></td>
<td>same time</td>
<td>9</td>
</tr>
<tr>
<td>2.4</td>
<td>Olympus EndoCapsule</td>
<td>10</td>
</tr>
<tr>
<td>2.5</td>
<td>Olympus Real-time viewer (taken from Olympus website)</td>
<td>10</td>
</tr>
<tr>
<td>2.6</td>
<td>Olympus EndoView video viewing software</td>
<td>11</td>
</tr>
<tr>
<td>2.7</td>
<td>Olympus EndoCapsule sensor array and data recorder (taken from Olympus</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>website)</td>
<td></td>
</tr>
<tr>
<td>2.8</td>
<td>Conceptual diagram of the future Olympus capsule guidance principle (taken</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>from Olympus New Release (Olympus News Release, 2004))</td>
<td></td>
</tr>
<tr>
<td>2.9</td>
<td>Norika capsule design (taken from RF System website (RF Systems, n.d.))</td>
<td>15</td>
</tr>
<tr>
<td>2.10</td>
<td>Sayaka capsule (taken from RF System website (RF Systems, n.d.))</td>
<td>15</td>
</tr>
<tr>
<td>2.11</td>
<td>IVP2 capsule (taken from IVP project website (Intracorporeal Videoprobe</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Project, n.d.))</td>
<td></td>
</tr>
<tr>
<td>2.12</td>
<td>IVP2 capsule steering system (taken from IVP project website (Intracorporeal</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Videoprobe Project, n.d.))</td>
<td></td>
</tr>
<tr>
<td>2.13</td>
<td>Gastrointestinal tract</td>
<td>17</td>
</tr>
<tr>
<td>2.14</td>
<td>Parts of the small intestine</td>
<td>18</td>
</tr>
<tr>
<td>3.1</td>
<td>Olympus image enhancement - Example of a raw image (left) and an enhanced</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>one (right). It is not clear whether such enhancements improve the quality</td>
<td></td>
</tr>
<tr>
<td></td>
<td>of the clinical diagnosis. Courtesy of Miguel Tavares Coimbra, Faculty of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sciences, University of Porto.</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>The structure of the topographic video segmentation methods</td>
<td>38</td>
</tr>
<tr>
<td>4.2</td>
<td>Gastric and Intestinal Transit Times</td>
<td>39</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

4.3 WCE images acquired from A) Mouth B) Stomach, C) Small Intestine, D) partially occluded Colon and E) completely occluded Colon; and below their respective equalized HS histograms. A visible shift in hue (vertical axis of the histogram) between the respective histograms is clearly visible. .......................................................... 44
4.4 LBP pattern with a circular neighbourhood .......................................................... 46
4.5 A grid of one hundred 16 × 16 macro blocks used for motion feature extraction. ................................................................... 49
4.6 A macro block of size sixteen and a search parameter of size seven ... 49
4.7 An example of a grid of motion vectors. It can be seen that most of the vectors point to the right. .................................................................. 50
4.8 A grid of unit length vector u pointing to the direction opposite to the origin (the middle point of the macro block grid). ......................... 51
4.9 Grid of 28 sub-images .................................................................................. 53
4.10 WCE images showing selected SubIRs. A-D Stomach; E-H Intestine. .................................................. 54
4.11 Three first principal components representing compressed histograms extracted from four different video regions .......................... 56
4.12 Compression zonal mask applied to 8 × 8 coefficient block. ..................... 58
4.13 Zonal mask in 3-D .................................................................................. 58
4.14 A figure illustrating the kNN algorithm - a classified vector with the k=5 closest neighbours pointed by arrows. ................................. 61
4.15 Two linearly separable data sets with separating hyperplane A and two hyperplanes H_1 and H_2 identifying the margin. ............... 63
4.16 Two linearly non-separable data sets with separating hyperplane A and two hyperplanes H_1 and H_2 identifying the margin. ............... 66
4.17 A simple segmentation method based on converging search ......................... 71
4.18 Sliding window video segmentation. E,S,I and C denote entrance, stomach, intestine and colon respectively. .................................................. 72
4.19 A figure showing the HMM for capsule video segmentation ......................... 73
4.20 A sequence of conditional class probabilities constituting input for the HMM ........................................................................ 74
4.21 A grid illustrating the search process of optimal C and γ parameters for a radial basis SVC involving LBP_3D_HS feature vector. ............... 77
4.22 A flowchart showing how discriminators 2, 4 and 6 choose the appropriate classifier for each frame from the classification sequence. ...... 82
4.23 Topographic video segmentation prototype deployed in Norfolk and Norwich University Hospital in Norwich. ........................................ 89
5.1 WCE images containing different manifestations of bleeding. .................. 92
5.2 Bleeding detection system ........................................................................ 96
5.3 Bleeding 3-D HSI histogram sliced along different intensity levels ............ 97
LIST OF FIGURES

5.4 Non-bleeding 3-D HSI histogram sliced along different intensity levels . 98
5.5 The flowchart illustrating how the dynamic updates of the histograms are performed .................................................. 99
5.6 ROC curves for $T_1 = 3, T_2 = -10$ and $\beta, \delta = 0$. Different curves represent different $\gamma$, along each curve increasing $\alpha$. .......................... 103
5.7 ROC curves for $T_1 = 3, T_2 = -10$ and $\gamma = 0.98, \alpha = 0.1$. Different curves represent different $\delta$, along each curve increasing $\beta$. .................. 104
5.8 $F$ measure curves for $T_1 = 3, T_2 = -10$ and $\beta, \delta = 0$. Different curves represent different $\gamma$. ......................................................... 104
5.9 $F$ measure curves for $T_1 = 3, T_2 = -10$ and $\gamma = 0.98, \alpha = 0.1$. Different curves represent different $\delta$. ......................................................... 105
5.10 Two images containing air bubbles with specular highlights. ............. 108
5.11 The shape of the MS diagram .......................................................... 109
5.12 Contrast enhancement and specularity detection: a,d) original and contrast enhanced images; c,f) their MS diagrams; and b,e) detected specular pixels .......................................................... 110
5.13 A grid illustrating the search process of optimal $C$ and $\gamma$ parameters for a radial basis SVC involving feature vector no. 8 from the table above. 114
5.14 Classification results - blood vs normal. (BvR) ................................. 115
5.15 Classification results - lesion vs normal (LvN) .................................... 117
5.16 Classification results - blood & lesion vs normal (LBvN) .................... 118
5.17 Graphic User Interface displaying the results of the blood detection ordered according to the likelihood of blood ........................................ 125
5.18 Graphic User Interface displaying one of the bleeding events in the video 126
5.19 Suspected Blood Indicator view in the Rapid Reader software ............ 127
A.1 Signed rank test results for naïve (1 pass) versus naïve (multiple passes) methods using LBP 3D feature vectors ........................................ 140
A.2 Signed rank test results for naïve (1 pass) versus naïve (multiple passes) methods using HS feature vectors ........................................ 141
A.3 Signed rank test results for naïve (1 pass) versus naïve (multiple passes) methods using DFT feature vectors ........................................ 142
A.4 Signed rank test results for naïve (multiple passes) versus sliding window (window sizes: 50, 200, 1200) methods using HS feature vectors 143
A.5 Signed rank test results for naïve (multiple passes) versus sliding window (window sizes: 50, 1200, 3000) methods using HS feature vectors 144
A.6 Signed rank test results for naïve (multiple passes) versus sliding window (window sizes: 50, 200, 1200) methods using LBP 3D feature vectors 145
A.7 Signed rank test results for naïve (multiple passes) versus sliding window (window sizes: 50, 1200, 3000) methods using LBP 3D feature vectors ........................................ 146
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.8</td>
<td>Signed rank test results for sliding window (window sizes: 50, 200, 1200) versus HMM methods using LBP 3D feature vectors</td>
<td>147</td>
</tr>
<tr>
<td>A.9</td>
<td>Signed rank test results for sliding window (window sizes: 50, 1200, 3000) versus HMM methods using LBP 3D feature vectors</td>
<td>148</td>
</tr>
<tr>
<td>A.10</td>
<td>Signed rank test results for sliding window (window sizes: 50, 200, 1200) versus HMM methods using HS feature vectors</td>
<td>149</td>
</tr>
<tr>
<td>A.11</td>
<td>Signed rank test results for sliding window (window sizes: 50, 1200, 3000) versus HMM methods using HS feature vectors</td>
<td>150</td>
</tr>
<tr>
<td>A.12</td>
<td>Signed rank test results for sliding window (window sizes: 50, 200, 1200) versus HMM methods using LBP 3D + HS feature vectors</td>
<td>151</td>
</tr>
<tr>
<td>A.13</td>
<td>Signed rank test results for sliding window (window sizes: 50, 1200, 3000) versus HMM methods using LBP 3D + HS feature vectors</td>
<td>152</td>
</tr>
<tr>
<td>A.14</td>
<td>Signed rank test results for Multivariate Gaussian versus SVC methods using HMM video segmentation and HS feature vectors</td>
<td>153</td>
</tr>
<tr>
<td>A.15</td>
<td>Signed rank test results for Multivariate Gaussian versus SVC methods using HMM video segmentation and LBP 3D feature vectors</td>
<td>154</td>
</tr>
<tr>
<td>A.16</td>
<td>Signed rank test results for Multivariate Gaussian versus SVC methods using HMM video segmentation and LBP 3D + HS feature vectors</td>
<td>155</td>
</tr>
<tr>
<td>A.17</td>
<td>Signed rank test results for Multivariate Gaussian versus SVC methods using HMM video segmentation and DFT feature vectors</td>
<td>156</td>
</tr>
<tr>
<td>A.18</td>
<td>Signed rank test results for Multivariate Gaussian versus SVC methods using HMM video segmentation and LBP 3D + DFT feature vectors</td>
<td>157</td>
</tr>
<tr>
<td>A.19</td>
<td>Signed rank test results for HS feature vectors extracted from entire images versus HS feature vectors extracted from entire images and SubIRs using SVC and HMM video segmentation</td>
<td>158</td>
</tr>
<tr>
<td>A.20</td>
<td>Signed rank test results for LBP 3D feature vectors extracted from entire images versus the same feature vectors extracted from entire images and SubIRs using SVC and HMM video segmentation</td>
<td>159</td>
</tr>
<tr>
<td>A.21</td>
<td>Signed rank test results for LBP 3D + HS feature vectors extracted from entire images versus the same feature vectors extracted from entire images and SubIRs using SVC and HMM video segmentation</td>
<td>160</td>
</tr>
<tr>
<td>A.22</td>
<td>Signed rank test results for LBP 3D + HS + DFT feature vectors extracted from entire images versus the same feature vectors extracted from entire images and SubIRs using SVC and HMM video segmentation</td>
<td>161</td>
</tr>
<tr>
<td>A.23</td>
<td>Signed rank test results for HS feature vectors versus LBP 3D feature vectors using SVC classifier and HMM video segmentation</td>
<td>162</td>
</tr>
<tr>
<td>A.24</td>
<td>Signed rank test results for DFT feature vectors versus HS feature vectors using SVC classifier and HMM video segmentation</td>
<td>163</td>
</tr>
<tr>
<td>A.25</td>
<td>Signed rank test results for HS feature vectors versus LBP 3D + HS feature vectors using SVC classifier and HMM video segmentation</td>
<td>164</td>
</tr>
</tbody>
</table>
A.26 Signed rank test results for LBP 3D feature vectors versus LBP 3D +
HS feature vectors using SVC classifier and HMM video segmentation . 165
A.27 Signed rank test results for LBP 3D + HS feature vectors versus LBP
3D + HS + DFT feature vectors using SVC classifier and HMM video
segmentation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 166
## List of Tables

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>The mean and standard deviation of the number of images with at least one SubIR for different GI parts.</td>
</tr>
<tr>
<td>4.2</td>
<td>Support vector machine kernels</td>
</tr>
<tr>
<td>4.3</td>
<td>List of entire image feature vectors tested in the first stage</td>
</tr>
<tr>
<td>4.4</td>
<td>List of entire image and SubIR feature vectors tested in the second stage</td>
</tr>
<tr>
<td>4.5</td>
<td>Percentage of correct classifications - First Stage</td>
</tr>
<tr>
<td>4.6</td>
<td>Percentage of correct classifications - Second Stage</td>
</tr>
<tr>
<td>4.7</td>
<td>List of discriminators tested in the first stage of the video segmentation experiment</td>
</tr>
<tr>
<td>4.8</td>
<td>Additional feature vectors tested in the second stage of the video segmentation experiment</td>
</tr>
<tr>
<td>4.9</td>
<td>The median and mean errors (in frames) of HMM discriminators tested in the first stage. Note: the entire video length is around 50,000 frames</td>
</tr>
<tr>
<td>4.10</td>
<td>The median and mean errors of HMM discriminators in frames. Note: entire video length is around 50,000 frames</td>
</tr>
<tr>
<td>5.1</td>
<td>List of feature vectors tested in the SVC experiment</td>
</tr>
<tr>
<td>5.2</td>
<td>Performance of SBI vs our method</td>
</tr>
</tbody>
</table>
Publications

The following are publications related to this work by the author:


- M. Mackiewicz, J. Berens, M. Fisher. Wireless Capsule Endoscopy colour video segmentation. Submitted to *IEEE Transactions on Medical Imaging*


## Glossary

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>Crohn’s Disease</td>
</tr>
<tr>
<td>DCT</td>
<td>discrete cosine transform</td>
</tr>
<tr>
<td>DFT</td>
<td>discrete Fourier transform</td>
</tr>
<tr>
<td>GI</td>
<td>Gastrointestinal</td>
</tr>
<tr>
<td>HMM</td>
<td>hidden Markov model</td>
</tr>
<tr>
<td>HSI</td>
<td>hue saturation intensity</td>
</tr>
<tr>
<td>kNN</td>
<td>k nearest neighbour</td>
</tr>
<tr>
<td>LBP</td>
<td>local binary pattern</td>
</tr>
<tr>
<td>MG</td>
<td>multivariate gaussian</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>SBI</td>
<td>suspected blood indicator</td>
</tr>
<tr>
<td>SubIR</td>
<td>sub-image region</td>
</tr>
<tr>
<td>SVC</td>
<td>support vector classifier</td>
</tr>
<tr>
<td>SVD</td>
<td>singular value decomposition</td>
</tr>
<tr>
<td>WCE</td>
<td>wireless capsule endoscope(y)</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Medical imaging (MI) refers to techniques used to acquire images from the areas of the human body that are not visible under normal conditions. It can be carried out for clinical purposes (medical procedures aiming to detect, diagnose or examine disease) or medical science (including the study of normal anatomy in an attempt to understand processes in humans). Many MI techniques depend on computers, whose aim is to aid physicians in understanding and interpreting the images.

1.1 Why does Wireless Capsule Endoscopy video analysis need computer assistance?

Wireless Capsule Endoscopy (WCE) is a recent and exciting area of MI, which involves recording images of the entire Gastrointestinal (GI) tract including the parts of the human body never before seen outside operative surgery. It is particularly suited for computer-assisted diagnosis, as it records a large quantity of data (mostly, but not exclusively images) from the human gut, which consequently requires a time-consuming visual assessment that can be carried out only by an experienced clinician. The duration of this assessment, which involves the scrutiny of a video comprising approximately
CHAPTER 1. INTRODUCTION

50,000 frames, varies between one to two hours. Thus, it can be seen that in terms of time requirement, the WCE is a very costly MI procedure. This opens a door for computers to aid the analysis of the WCE footage, by reducing the time required to reach the diagnosis and thus the cost of the procedure, making it a more affordable technique. This view is supported by the leading endoscopists in the United Kingdom:

"The cumbersome analysis of WCE images has been the major faction in preventing the spread of WCE to become a routine service in every DGH."

Dr Jonathan Green, Secretary of the Endoscopy Committee of the British Society of Gastroenterologists.

"The key constraint for uptake of WCE is the time taken to read and report the examination. Increased automation of reading of WCE will have a dramatic effect on the uptake of this important endoscopic technique, and to the costs of providing the service." Dr Roland Valori, National Endoscopy Lead.

Another aim of computer-assisted WCE video analysis can be considered in terms of improvement of the clinicians diagnosis. Here, the computer assumes the role of an expert system able to give a second opinion on the state of the patient.

1.2 Main areas of research contained in the thesis

The main areas of research that have been studied within the framework of this thesis are topographic video segmentation and bleeding detection. The former considers the problem of segmenting the capsule video into meaningful parts such as mouth, oesophagus, small intestine and colon. Performing such segmentation automatically creates certain advantages for clinicians, which will be discussed later in this thesis. We investigated a range of computer vision and image processing techniques which provided excellent
results. These techniques include various image feature extraction methods that enable image content to be modelled in mathematical terms. The choice of the right features is probably the most important issue in the entire topographic video segmentation task. We want to have a set of features that can be easily and quickly computed, but are capable of robustly discriminating between the several target classes. The work described in this thesis uses content features, which are the features derived directly from an image. Such features are based on color, texture and motion information. In this work we use image features extracted from compressed Hue-Saturation (HS) and Local Binary Patterns (LBP) histograms. The motion features were extracted from motion vectors calculated from a grid of corresponding square blocks in pairs of consecutive frames. These were further analysed in the context of the neighbouring frames by transformation into the frequency domain using the Discrete Fourier Transform (DFT). A large number of features were produced and consequently different compression methods were investigated, including Discrete Cosine Transform (DCT) and Principal Component Analysis (PCA).

The next stage of video segmentation involves classification of image features into certain previously mentioned anatomical classes. Here, we investigate different pattern analysis and recognition algorithms, which enable correct classification of extracted features. The techniques used here included various linear and non-linear classifiers: Multivariate Gaussian, kNN and Support Vector Classifier (SVC).

In the capsule video segmentation final stage, we use the sequence of classified features to perform the actual video segmentation, i.e. label the transition points between anatomical regions. Here, we investigate different segmentation methods which include naïve search, the sliding window method and Hidden Markov Model (HMM).

Another area of research addressed in this project is bleeding detection. In this scenario, the aim is to automatically find those frames in the video that contain signs of bleeding. Although proprietary software of the most popular WCE manufacturer
(Given Imaging Ltd, n.d.) includes such a functionality (Suspected Blood Indicator (SBI)), clinical reports regarding the SBI are not positive i.e. they report low sensitivity and specificity of bleeding detection (Signorelli et al., 2005; D’Halluin et al., 2005).

In this work, we use a range of computer vision and pattern recognition techniques to mitigate these problems and improve the detection accuracy. In order to achieve this, we built an adaptive model of blood and non-blood colour distributions. This model is implemented using 3-D Hue Saturation Intensity (HSI) histograms. A model that uses histograms can be easily updated, according to the changing colour contents of the video. Image pixels whose HSI colour values fall above certain pre-calculated thresholds are considered suspicious and are used in a region growing process to form a candidate bleeding region. The next step involves extraction of low level features (3-D HSI and LBP histograms) from candidate blood regions and their surrounding neighbourhood. Moreover, the region is tested to find whether it contains specular reflectances that can suggest a presence of the air bubble, and which is indicative of a possible false detection. Extracted HSI and LBP features are then used to classify the candidate blood region into bleeding, lesion/abnormality and non-blood classes. To perform this task, a non-linear Support Vector Classifier was chosen, as it provided the best results in the earlier topographic video segmentation experiments.

Other products of this research are the prototypes of both topographic video segmentation and the bleeding detection tools. These have been deployed in the Norfolk and Norwich University Hospital (NNUH) for evaluation.

1.3 Thesis organisation

The two research areas investigated in this thesis can be considered as the first natural steps in the problem of computer-assisted analysis of capsule videos. The topographic video segmentation problem considers the WCE video at the highest level in a hierarchy
of video content, as expressed by the ordered sequence of anatomical events, which are the same for every subject (with an exception of extremely rare abnormal post-surgical cases). The second area of research (bleeding detection) focuses on a lower level of the hierarchy, which maps onto events i.e. there can by multiple bleeding events in any of the anatomical regions. What is even more relevant is the fact that the visual appearance of a bleeding event can depend on the anatomical region it is located in. Therefore, it is important for the bleeding detection algorithm to be aware of the current frame location with respect to the anatomy. The algorithm proposed in this thesis uses this observation - it was found that locating the pylorus (the valve between the stomach and the small intestine) is crucial to the issue of bleeding detection. Thus, confirming bleeding detection research as a natural continuation of the work on topographic video segmentation.

1.3.1 Thesis outline

The thesis is organised in the following manner. In Chapter 2, WCE technology and its clinical use are described. Moreover, we also discuss in details the motivation and background for this research in more detail. Chapter 3 comprises a literature review of computer vision and image processing techniques and algorithms, which were used by other researchers in the analysis of the WCE footage. Chapters 4 and 5 contain a description of the main body of research carried out within the framework of this project. The former discusses the advances in the area of topographic video segmentation and the latter describes our contributions in the field of bleeding detection. In Chapter 6, we conclude and discuss the possibilities of future work.
Chapter 2

WCE technology and its clinical importance

This chapter provides a comprehensive description of the Wireless Capsule Endoscope technology. The products of different manufacturers are described, with an emphasis on the first device introduced by Given Imaging Ltd. (Given Imaging Ltd, n.d.), which has been the most widely researched and is also used in the experiments described within this thesis. A review of the medical publications that confirm the importance of the technology and its clinical use can be found in Section 2.2. Section 2.3 continuous to develop the motivation for computer vision tools in WCE and finally some conclusions regarding the matters discussed in this chapter are drawn in Section 2.4.

2.1 Technology

The first wireless capsule endoscope was launched in 2001 by Given Imaging Ltd, and reported in an article in "Nature" (Iddan et al., 2000). Since the device received FDA (American Food & Drug Administration) clearance in August 2001, over 400,000 examinations globally have been conducted. The 11mm x 26mm M2A capsule (later
rebranded PillCam SB (SB stands for small bowel)) (see Figures 2.1 and 2.2) is propelled passively, hence having been swallowed, it is propelled through the food tract by normal peristaltic movement of the human gastrointestinal (GI) system, usually reaching the colon, before being expelled naturally from the body. One end of the capsule contains an optical dome with six white Light Emitting Diodes (LEDs) and a CMOS camera that captures 2 images (circular shape from a square of $256 \times 256$ pixels) a second. These images are compressed using JPEG and relayed via a transmitter using a radio frequency signal (approximately 432 MHz) to an array of aerials, which are attached to the patient’s body, from where they are transferred over the wires to a data-recorder worn by the patient on a belt. The sensor array allows for continues triangulation of the position of the capsule inside the body of the patient so that the trajectory of the capsule passage can be later displayed on the workstation monitor. The accuracy of the capsule location provided by this method was reported to be $\pm 3$ cm (Ravens and Swain, 2002). After 8 hours (the capsule two silver-oxide batteries lifetime), the data-recorder is removed and the image data uploaded to a workstation for viewing and analysis. The upload process originally took two to three hours using early versions of the Given Imaging Rapid Reader (RR) software, but a more recent version of this application has reduced this to around 30 minutes. The stored data consists of $\sim 50,000$ images, and is viewed as a video sequence using software provided by the manufacturers (RR) (see Figure 2.3)).

The clinical procedure is a simple and painless process, which is one of the factors why the patients prefer WCE to conventional endoscopy methods (Melmed and Lo, 2005). Before the procedure begins, patients must fast overnight. Moreover, more recent studies suggest that results may be improved by bowel preparation i.e. ingesting a drug that shortens food and capsule transit times through the GI track (Dai et al., 2005). The exam begins by the attachment of the antennas to the patient’s chest, which are then connected to the data-recorder worn on a belt. The capsule starts acquiring images and
transmitting them immediately after it is removed from its magnetic holder. After a brief capsule test by the physician (less than one minute), it should be immediately ingested by the patient. Water and food intake can begin after 2 and 4 hours respectively. Patients are asked to monitor a blinking diode on the belt pack, which indicates the quality of the signal reception. The patient should not exercise during the procedure, and avoid any powerful electromagnetic field source.
In October 2005, Olympus launched a competitor system called EndoCapsule (see Figure 2.4) in Europe. Their device, acquires images at the same rate and at the same resolution as the Given Imaging PillCam. The difference lies in the use of a different imaging technology - CCD, which the manufacturers claim is of higher quality (Fuyono, 2005). Another feature of the EndoCapsule is the Automatic Brightness Control (ABC), which was applied from the traditional endoscope technology, of which Olympus are the world leaders. ABC provides an automatic illumination adjustment as the conditions in the GI track vary. EndoCapsule also features a real-time viewing device (see Figure 2.5), which allows the physician to watch the capsule examination in real time. Moreover, the EndoView software also features image pre-processing capabilities, which result in a high image quality.
In December 2004, FDA approved a second type of capsule developed by Given Imaging - the PillCam ESO, which allows the evaluation of oesophageal disease. The
motivation for developing this capsule lay in the rapid capsule transit through the oesophagus. Here the frame rate of only two frames per second (of the original PillCam SB) did not provide enough data for a thorough examination of this GI region to be undertaken. This opinion was also reported in (Neu et al., 2003), where the authors concluded that esophageal capsule endoscopy could not be achieved without further technical developments. The response to this demand materialised in the development of the PillCam ESO which has the higher frame rate and CMOS cameras positioned at both ends of the capsule. This capsule acquires and transmits seven frames per second from each camera, giving a total of 14 frames per second (Mishkin et al., 2006). Due to the increased frame rate, the capsule battery life is only 20 minutes, which is ample time for the capsule to visualise the entire oesophagus.
In October 2006, Given Imaging received the CE Mark to market a third capsule - the PillCam COLON throughout the European Union. This device was developed as a diagnostic test to visualize the colon. The capsule measures 11 mm by 31 mm - slightly larger than the previous two capsules. Similar to PillCam ESO, the capsule has cameras at both ends. These capture 4 images a second for up to 10 hours. A new feature in Given Imaging capsules is an automatic lighting control. Since the lumen of the colon is wider than the small bowel and also highly compartmentalised, the PillCam COLON capsule optics captures more than twice the coverage area and depth of field of the PillCam SB capsule. After switching on the capsule and within a few minutes of image transmission, the device enters a delay mode. This lasts approximately 2 hours, after which it "wakes" up to resume the image transmission. This saves the battery power during the transit of the capsule through the earlier parts of the gut, allowing longer
transmission from the relevant GI region - the colon.

2.1.1 Future development directions

All active endoscopes (those whose movement can be controlled) produced so far are wired i.e. they transmit the images from the camera to the display over a wire. The capsules described above are all passive, which means they are propelled by peristalsis and their movement cannot be controlled. An active propulsion system would offer obvious benefits. It is not surprising, therefore, that it is the focus of intensive research. In (Sendoh et al., 2003), the authors propose a new approach of propelling the capsule using a magnetic accurator. The device uses a permanent magnet in the capsule and an external rotational magnetic field to control the capsule movement. The authors report a preliminary experiment in which they used a dummy capsule without any endoscopic functions.

Olympus is also working on the development of a new generation capsule endoscope, which features magnetic propulsion (Olympus News Release, 2004) (see Figure 2.8). Apart from the novel propulsion and guidance system, the capsule designers aim to provide a drug delivery system, which would administer drugs directly to an affected area; a body fluid sampling system, taking body fluid extracts for diagnosis and analysis; and also an ultrasound scan capability.

RF System Lab Company (RF Systems, n.d.) has also announced the intention of producing a Norika capsule with a magnetic field based propulsion (see Figure 2.9). In December 2005, they also announced the design of the new Sayaka capsule, which would have a lens on the lateral surface of the capsule instead of the front as in the capsules of the competitors (see Figure 2.10). The inventors claim that such a design would obtain clear-cut images of the gastrointestinal wall while the capsule spins in the GI tract with the stepping angle of 7.5°. Sayaka acquires images at a rate of 30 frames per second, which generates \( \sim 870,000 \) over an eight hour period of operation. The
images are later combined together in a process called mosaicing, which produces an image (map) of the GI tract walls.

Further applications of magnetic fields are presented by (Lenaerts and Puers, 2006), where the authors propose to use an inductive link to power a wireless capsule endoscope. The capsule, known as the Intracorporeal Videoprobe (IVP2, see Figure 2.11) (Arena et al., 2005) is induction-powered and equipped with a tilting image sensor on a motorised plate (see Figure 2.12) and a telemetric datalink. In (Turgis and Puers, 2004), the authors describe the video compression method of this system. According to (Innovative imaging probes for endoscopy, 2005), the capsule is still a long way from becoming a commercial product.

![Conceptual diagram of the future Olympus capsule guidance principle (taken from Olympus New Release (Olympus News Release, 2004))](image)

**Figure 2.8:** Conceptual diagram of the future Olympus capsule guidance principle (taken from Olympus New Release (Olympus News Release, 2004))

### 2.2 WCE clinical importance

The gastrointestinal (GI) tract (see Figure 2.13) consists of the oesophagus, stomach and duodenum (upper GI tract), the jejunum, ileum (small bowel), colon and rectum.

Fibre optic gastrointestinal endoscopy, introduced in the early 1970s enabled effective diagnosis and biopsy of disease in the lumen of the stomach and duodenum (gastro-duodenoscopy OGD) and shortly afterwards the colon and rectum (colonoscopy). In
the 1980s, videoendoscopy using microchip cameras improved image resolution and the ease of use of the equipment.

These innovations enabled clinicians to diagnose a large number of GI pathologies which occur mostly in the stomach, duodenum and colon. Examination of the remainder of the small intestine using a conventional endoscopy (the jejunum and ileum) was until very recently limited only to the first few centimeters of jejunum and last few centimeters of the terminal ileum. However, this part of the GI tract can be the site
of "obscure" bleeding, inflammation and a rare location for tumours. There exist non-standard endoscopy methods to investigate this part of the small bowel. However, they use a flexible tube to propel the endoscope along the bowel, which given the length of the investigated gut (up to 4.5) and its shape with many loops, rendered the procedure very difficult and uncomfortable to patients. Other means of imaging the small bowel, such as computerised tomography (CT) or magnetic resonance (MR) axial imaging cannot provide a direct view of the tissue. Hence, there was a clinical need to provide a tool which would enable clinicians to diagnose the jejunum and the ileum more efficiently.

The wireless capsule endoscope (WCE) developed by Given Imaging Ltd was an answer to this demand, which for the first time allowed an endoscope to record the high-resolution images of its full passage through the GI tract including the parts of the human body never before seen outside operative surgery.
CHAPTER 2. WCE TECHNOLOGY AND ITS CLINICAL IMPORTANCE

Following many clinical studies (Gay et al., 2004; Leighton et al., 2006; Selby, 2004; Viazis et al., 2005) into the efficacy of WCE in clinical practise, the superior imaging capabilities of WCE above contrast imaging and CT imaging became clear. These studies concentrated on patients with "obscure" GI bleeding; those who had lost blood into the lumen with no cause found at OGD or colonoscopy. Consequently, WCE is now established as the first line diagnostic test for patients with this condition, and received approval as such from the National Institute for Clinical Excellence (NICE) in the UK in 2004.

WCE also received attention with regard to the diagnosis and assessment of patients with suspected or known Crohn’s disease (CD) (Bona et al., 2006; Dubcenco et al., 2005; Pennazio, 2004b,a; Swain, 2005), a chronic inflammatory disease affecting any part of the GI tract, but most often localised in the terminal ileum. This disease is of increasing prevalence in the developed world (Bernstein, 2006), and causes significant morbidity to patients. It is difficult to diagnose, partly due in the lack of sensitivity of
contrast radiology to minor mucosal pathologies. WCE offers significant improvements to the diagnosis of CD due to its greater sensitivity, and also allows follow-up examinations without exposing the patient to the hazards of ionising radiation. The capsule has also been found useful in the diagnosis of another form of small bowel inflammation - Celiac disease (Culliford et al., 2005).

Cancer of the small bowel is rare but often has a high mortality rate since it tends to be diagnosed late. A number of clinical studies (Cobrin et al., 2006; van Tuyl et al., 2006; Urbain et al., 2006) have found WCE as a sensitive tool in detecting cancerous conditions. Therefore, if employed with this group of patients, capsule endoscopy is expected to offer a clinical benefit in terms of cancer survival rate. Peutz-Jegher disease is a hereditary condition that involves the formation of small bowel polyps with a potential to become malignant, especially when they reach a large size. In (Brown et al., 2006), the authors suggest that wireless capsule endoscopy should become the investigation of choice in those undergoing follow-up in interval surveillance for this
condition.

Patients with obscure diarrhoea abdominal pain and possible functional bowel disease have also been examined in series of studies to assess potential diagnostic benefits of WCE, but in this heterogeneous group no clear benefit has been demonstrated to date (Fry et al., 2006). The capsule has also proven useful in studies concerning the impact of drugs on the gastrointestinal tract (Qureshi, 2004). Moreover, children can benefit from the device as well as adults (Argüellas-Arias et al., 2004).

The development of PillCam ESO with a faster frame acquisition rate has allowed the non-invasive imaging of oesophageal conditions in selected groups of patients. This safe and acceptable-to-patients form of WCE may represent an alternative to conventional upper endoscopy, and give a clinical benefit in the assessment of endoscopic signs of esophageal varices and portal hypertension (Eisen et al., 2006; Eliakim et al., 2004; Caunedo-Alvarez and Herreras-Gutierrez, 2006). Early studies from the PillCam COLON (Eliakim et al., 2006; Schoofs et al., 2006) launched for clinical use by Given Imaging in 2006 show promising accuracy compared with colonoscopy and suggest further clinical trials to establish the efficacy and appropriate role for this application of capsule endoscopy.

It is believed by some researchers (Iddan et al., 2000) that WCE approximates ‘physiological endoscopy’ as it is more closely aligned with human physiology than conventional endoscopy, where sedation and air insufflation are required to provide a clear passage for the conventional wired scope. These factors result in the distortion of the natural appearance of the viewed area in traditional endoscopy and colonoscopy.

2.3 Clinical need for better computational tools

WCE exam viewing times vary from 45-90 minutes, and depends on the clinician’s experience, the complexity of the case and the small bowel transit time range. The Sus-
expected Blood Indicator (SBI) (an automated tool designed to detect images containing bleeding in the video, for more information see Section 5.2) in Given Imaging Rapid Reader software, may be helpful, but according to (Signorelli et al., 2005) does not replace the need for a full video viewing and should be used only as a complementary tool. An important drawback to the Given Imaging system is the time spent by a clinician identifying the landmarks of pylorus and ileocaecal valve, which allows the clinician to focus on the small bowel section of the video. This constitutes a time burden on clinicians, particularly in cases where these two landmarks are difficult to find. In (Lai et al., 2006), some controversy regarding inter-observer agreement between those of varying experience and training have been reported. In (Ravens and Swain, 2002), the authors comment that with the expected reduction in capsule prices, the time needed by a clinician to analyse the exam may soon become the most expensive part of the procedure. Thus, a reduction of this time would be a major benefit, provided the quality of the diagnostic report was not reduced. The existing systems have user-friendly viewing interfaces, but with few exceptions lack automated tools that would highlight places of interest. Such tools could not only shorten the exam viewing time, but also improve the quality of patient’s diagnosis by drawing attention to possible pathology, which could have been missed by the clinician among many thousands of normal frames. Incidentally, the manufacturers of the capsule try to reduce the video viewing time using additional viewing controls e.g. double and quad views in the Rapid Reader and EndoView software packages (see Figures 2.3 and 2.6); jog/shuttle control device (used instead of a mouse as a video control, manufactured by Control Design Inc. Windham, New Hampshire, USA) (Seitz et al., 2005).

Another important issue is data storage, which should be considered in terms of security, stability and volume management. The capsule exam videos may need re-examination after a delayed interval. Loss of data must be avoided at all cost since it would represent a clinical governance risk. The storage of all electronic patient data
must also conform with national guidelines for electronic patient records. WCE reporting software could be integrated into existing endoscopic reporting packages, allowing for easier audit of departmental practices.

2.4 Conclusions

Since its advent in 2001 Wireless Capsule Endoscopy technology has rapidly grown - resulting in a number of commercial products. The current technology can be regarded as the first generation of WCE. It can be characterised by relatively low frame rate, limited battery life and consequently the necessity of trading-off one for another. All the capsules on the market are passive devices, in contrast to the second generation devices which are likely to involve EM field propulsion. These devices will also feature a remote power source, allowing higher resolutions and frame rates. Moreover, other useful functions such as biopsy, drug delivery and telemetric measurement capabilities might be added.

As far as clinical importance is concerned, small bowel WCE has proven to be a very successful diagnostic tool in many patients with clinically significant conditions often outperforming existing techniques. The procedure is safe and does not involve any radiation risk; it is acceptable to patients, requires little nursing supervision, as is usually an ambulatory outpatient procedure. Development of WCE in the oesophagus and particularly the colon is at a relatively early stage of clinical assessment. However, it is likely that in the future, these methods could replace some conventional diagnostic endoscopy procedures.

The cost of the procedure remains high, which is partly due to the amount of clinician’s time required for reporting each patient. Hence, there is a clear need for automated tools capable of detecting certain events in WCE videos. The next chapter reviews the relevant work published in recent years in this field.
Chapter 3

Computer Vision in WCE video analysis - literature review

This chapter provides an extensive literature review of computer vision techniques in WCE video analysis. Those topics in which this thesis contributes: *topographic video segmentation* and *bleeding detection* are covered separately in appropriate sections in Chapters 4 and 5. Here we consider other WCE video analysis subjects such as: *abnormality detection*, *intestinal fluid detection*, *detection of intestinal contractions*, *capsule speed estimation*, *adaptive video viewing speed adjustment* and *WCE image enhancement*. Moreover, in Section 3.2, we review several research topics with regard to the use of image processing and computer vision in conventional endoscopy and colonoscopy. Finally, the issues raised in this Chapter are summarised in Section 3.4.

3.1 Application of computer vision in WCE

When the work on this thesis began in 2003, there were no publications about capsule endoscopy in the computer vision and image processing literature. Since then, wireless capsule endoscopy (WCE) has received significant attention from these communities.
Although, the use of image processing in WCE video analysis is still in its infancy, a significant number of papers have already been published. The applications of computer vision in capsule image analysis can be divided into four categories. The first category, which has probably received the largest attention judging from the number of papers published, considers the topographic segmentation of WCE video into meaningful parts such as mouth, oesophagus, stomach, small intestine and colon. This thesis brings the major contribution in this area, hence a comprehensive review of other authors’ work in this field is given in Section 4.3. The second category involves the detection of clinically significant video events (both abnormal and normal). Examples include physical abnormality (see Section 3.1.2), intestinal fluids (see Section 3.1.1), intestinal contractions (see Section 3.1.4) and capsule retention (see Section 3.1.3). This category also includes bleeding, an area in which we contribute in this thesis; and hence the description of the relevant literature will be given in Section 5.2. The third category considers video analysis with respect to changes in consecutive frames, in an attempt to adaptively adjust the video viewing speed (see Section 3.1.5) and hence achieve a reduction in the viewing time. The final area of research is the use of image processing techniques focused on the quality enhancement of raw images captured by the capsule (see Section 3.1.6).

### 3.1.1 Intestinal fluid detection

In (Vilarino et al., 2006), the authors present an algorithm which detects areas in the WCE video comprising images completely obscured by intestinal fluids. Early detection of such regions is highly beneficial since they can be removed from the sequence, before it is presented to the clinician, resulting in a shortening of the reviewing time. Intestinal fluids appear as yellowish to brownish semi-opaque turbid liquids often containing air-bubbles as well as other artifacts. The authors point out that the most relevant feature of the intestinal fluids is the presence of small bubbles of different sizes
and quasi-circular shapes. The algorithm is based on texture analysis performed using Gabor filter banks. In order to construct a filter bank, the authors used four different directions oriented at $0^\circ$, $45^\circ$, $90^\circ$, $135^\circ$, arranged of four gaussian scales (sigma values of 1, 2, 4 and 8), resulting in 16 filters in the bank. Frames that contained bubbles detected in more than 50% of the useful visualisation area were considered not valid for clinician analysis. The authors tested their algorithm on ten WCE videos in which the reduction in number of frames varied from 12 to 46% (mean 23%). They plan to expand this work to consider postprandial cases, where the texture patterns are more irregular.

### 3.1.2 Abnormality detection

In (Boulougoura et al., 2004), Boulougoura et al. describe an intelligent system, which they claim is capable of discriminating between normal and abnormal tissue in WCE images. They use 54 feature vector elements, incorporating nine statistical measures (standard deviation, variance, skew, kurtosis, entropy, energy, inverse different moment, contrast and covariance), calculated from histograms of six channels ($R,G,B,H,S,V$). The images are classified using an advanced neural network scheme containing the fusion of multiple classifiers dedicated to specific feature parameters. The authors report a detection accuracy of 100%. However, this result was evaluated using only 73 capsule images (33 abnormal and 38 normal), split into the training set (23 abnormal, 25 normal) and the test set (10 abnormal, 13 normal). The size of data used in this study is insufficient to draw conclusions as to whether the system can be used in a working application.

In (Coimbra, Campos and Cunha, 2006a), the authors measure the usefulness (classification potential, inter-coefficient redundancy, etc.) of MPEG-7 visual descriptors for detecting a variety of events (bleeding, ulcers and polyps), pointing out the superior performance of the Scalable Color and Homogenous Texture descriptors. The authors also conclude that the actual classification results are still far from ideal and hence fur-
ther research into more complex classification methods as well as better features are necessary.

3.1.3 Capsule speed estimation/retention detection

In (Szczypinski et al., 2004a,b), the authors localise the areas in the GI tract that might be affected by Crohn’s Disease (CD). In this work a Model of Deformable Rings (MDR) is used to locate the areas where the capsule moves more slowly or stops. This, according to some researchers (Tang et al., 2003) may signify the appearance of CD. The model aims to determine the movement of a tube-like surface (GI tract) by comparing adjacent video frames with regard to the displacement of its distinctive portions. MDR calculates a 2D map of the internal surface and provides an estimate of the capsule velocity. The map can be used as a quick reference, supporting identification of the segments of the GI tract and according to the authors claim, may be useful in identification of large scale pathologies.

3.1.4 Detection of intestinal contractions

Intestinal contractions, which are of some relevance to clinicians constitute only around 1% of the WCE video. In (Vilarinao et al., 2005) Vilarino et al. use ROC curves with ensembles of classifiers to detect these contractions based on 34 low-level image features from 9 consecutive frames including: mean intensity; hole size; global contrast; correlations between three previous sequences; and the corresponding sequences averaged across the objects for the class "contractions" and the variance of intensity. In (Spyridonos et al., 2005), the authors introduce a two stage contraction detection algorithm based on a Support Vector Classifier. Patterns of intestinal motility are encoded using a number of textural and morphological features, that include: 1st order statistics (mean, standard deviation, skew, kurtosis estimated from the image histogram); 2nd
order statistics (energy, entropy, inertia, local homogeneity, cluster shade and cluster prominence); Local Binary Pattern histograms (radius = 2, number of points = 16); and morphological features of the intestinal lumen (blob area, blob shape (solidity), blob sharpness and blob deepness). The authors report 73.5% sensitivity, 98.8% specificity and a false alarm ratio of 60%. The above figures were calculated from the test set comprising of 6 capsule videos.

3.1.5 Adaptive viewing speed adjustment

The main motivation for applying computer vision techniques to WCE video analysis is the potential improvement gained by reducing the overall time needed to review the data, by alerting the expert to clinically significant video frames. This may be achieved not only by automatic detection of events or segmenting the video into some meaningful parts, but also by adjusting the replay speed (number of frames displayed per second).

In (Hai et al., 2006), the authors propose such a method of handling the frame rate in a capsule image sequence. In their solution, instead of letting the clinician adjust the frame rate manually, video speed is adjusted by an algorithm, which plays the video at high speed in stable regions and at slower speed where significant changes between frames occur, signifying the possibility of pathologies. The authors divide each frame into 64 blocks and measure the similarity of colours between respective blocks in consecutive frames. \( RGB \) histograms quantised to \( 16^3 \) bins are used to describe each image block. The distance between local histograms is computed using the \( L_1 \) norm, formally:

\[
D_{blk}(i) = \sum_{k=1}^{N_{bins}} (|H_{R,k}^n - H_{R,k}^{n+1}| + |H_{G,k}^n - H_{G,k}^{n+1}| + |H_{B,k}^n - H_{B,k}^{n+1}|) \tag{3.1}
\]

which is later used to calculate the similarity between two frames:
\[ Sim(n) = \frac{1}{N_{blocks}} \sum_{i=1}^{N_{blocks}} sim\_block(i) \] (3.2)

where
\[
\begin{align*}
\text{sim\_block}(i) &= 1 \quad \text{if} \quad D_{blk}(i) > Thresh_{block} \\
\text{sim\_block}(i) &= 0 \quad \text{otherwise}
\end{align*}
\] (3.3)

Moreover, the maximum and minimum of distances between blocks are collected:
\[ D_{max}(n) = \max_i(D_{blk}(i)), \quad D_{min}(n) = \min_i(D_{blk}(i)) \]. In addition to colour features, the algorithm estimates motion displacement by extracting features using the KLT algorithm (Shi and Tomasi, 1994), tracking them using Newton-Raphson iterations. Then, \( Motion(n) \) is defined as the maximum of all displacements and together with \( Sim(n) \) are used to classify each frame by the decision tree into four states: 1) capsule and small intestine are stationary; 2) movement of the small intestine is small; 3) the small intestine has larger movements and finally 4) the small intestine has abrupt changes.

From the sequence of states, the delay time between consecutive frames is calculated using parametric functions, which take into account: the state to which the frame belongs, \( Motion(n) \), \( Sim(n) \), the skill of the clinician and the hardware limitations. The authors conclude that using their method the viewing time may be reduced from 2h to around 30 minutes without loss of information.

The software supplied by both Given Imaging (Rapid Reader) and Olympus (EndoView) also include play speed control. Unfortunately, the details of these algorithms remain unknown. Moreover, in the most recent version of Given’s Rapid Reader Application (V4.1), the clinician is given an option of watching a video in either "Normal Mode" or in the "Quick View Mode". Although the "Quick View" mechanism is not precisely explained in the documentation, we noticed that it uses an approach similar to that described above to reduce the viewing time of the video. It must be added though, that the "Quick View" mode skips some frames, displaying only the most suspicious (at least to the algorithm that is used by Given Imaging), which makes it different to the
algorithms described above and puts it closer to the category of algorithms aiming to
detect abnormalities, as described in Section 3.1.2.

The obvious conclusion regarding these methods must be that they are highly sub-
jective. All research on this topic has to include particularly extensive clinical evalua-
tions.

### 3.1.6 Image quality enhancement

Image quality enhancement is an area of image processing application that is also very
subjective and difficult to evaluate. Apart from standard noise reduction methods, it
might be possible to visually enhance the image so that relevant events are easier to
spot. The first commercial example (see Chapter 2 Olympus EndoView viewing soft-
ware), probably uses some sort of contrast and texture enhancement algorithms before
displaying the captured images. Our knowledge of this method is limited as the precise
algorithm has not been revealed. Although the images indeed look more appealing than
the original images and the images produced by the Rapid Reader software, we cannot
be sure whether this method increases the diagnostic yield of the capsule exam or just
creates misleading visual artifacts (see Figure 3.1). This method as well as any other
similar methods that will surely develop in future require deep and unbiased clinical
trials.

### 3.2 Application of Computer Vision in conventional en-
doscopy

Computer vision algorithms have been applied to conventional wired endoscopy long
before the development of their wireless counterpart. Hence, the area is much more
developed and consequently the literature is much richer. The list of issues touched
upon in this section, by no means should be considered full, but rather a selection of examples that may turn out relevant for future WCE research. The topics discussed include mostly studies related to colonoscopy (see Section 3.2.1) and endoscopy (see Section 3.2.2) image analysis as well as conventional endoscopy image retrieval system (see Section 3.3).

### 3.2.1 Colonoscopy image analysis

Krishnan et al. (Krishnan et al., 1998) presented a method of computerized detection of abnormality from colonoscopy images. An abnormality is detected from the curvature change of a haustra creases contour. Curvature of each contour is calculated, followed by the detection of the zero-crossings. The abnormality is identified when there is a contour segment between two zero-crossings having the opposite curvature signs to those of the two neighbouring contour segments. This proposed method can detect the presence of abnormalities such as polyps or tumours.
CHAPTER 3. LITERATURE REVIEW

In (Wang et al., 2001), a two stage texture based method of classifying colon tissue is presented. In the first step, the texture information is extracted. The texture features are derived from Local Binary Patterns (LBP), obtained from a $3 \times 3$ neighbourhood. A log-likelihood-ratio, the G statistic, which is a modification from Kullback’s criterion, is used as a pseudo-metric criterion. If an image contains different features preliminary segmentation is carried out and a Self-Organizing Map (SOM) is used in a classification step. The results suggest that the method can be used in unsupervised colonoscopy image classification.

In (Maroulis et al., 2001) a colorectal lesion detector is described. The system utilises second-order statistical features derived from the co-occurrence matrix (angular second moment, inverse different moment, correlation and entropy), which is calculated for four angles from the wavelet transformation of each colonoscopy image, to discriminate between normal and abnormal tissue. Artificial neural networks are used to classify these features. The detection accuracy of the system has been estimated at 95%, which allows its use as a supplementary diagnostic tool for detection of colorectal lesions such as pre-cancerous polyps.

In (An et al., 2005), the authors propose a method of detecting out-of-focus frames in the colonoscopy sequence. These frames constitute on average 37% (maximum 60%) of the colonoscopy video and are considered non-informative. Hence, they should be removed before the further processing either by a human expert or a computer-aided system. The informative and non-informative frames are classified using clustering based on features obtained from the Discrete Fourier Transform (DFT) and texture analysis. These features are extracted from the frequency spectrum image using the gray level co-occurrence matrix (GLCM) and include the following: entropy, contrast, correlation, homogeneity, dissimilarity, angular second moment and energy. The authors perform two step clustering using a K-means algorithm. The first step attempts to divide the data into 3 clusters: informative, non-informative and ambiguous, whereas
in the second step, the class *ambiguous* is further subdivided into *informative* and *non-informative*. This scheme was reported to be more successful than simpler and more intuitive one step, two cluster method, achieving an overall accuracy of 96%.

Hidović and Claridge (Hidovic and Claridge, 2005) present a novel optical imaging method capable of extracting parameters depicting histological quantities of the colon. Their approach uses a physics based model of light interaction with tissue, in this case colon, which is modelled with three layers: mucosa, submucosa and muscle layer. These layers have optical properties, which are defined by molar concentration and absorption coefficients of haemoglobin, the size and density of collagen fibres, the thickness of the layer and the refractive indexes of collagen and the medium. The histologically plausible ranges of these parameters are known and can be used to computationally create a cross-reference between the histological quantities and the associated spectra. The findings in the paper suggest that it should be possible to compute histological quantities from the colon multi-spectral images and hence significantly improve the diagnosis of the colon tissue.

### 3.2.2 Endoscopy image analysis

In (Majewski and Jedruch, 2005), the authors propose a computer-based system as a way to classifying malignant versus benign tumours in the gullet. The classification scheme uses features based on the structure of edges (extracted using a Water-Filling algorithm introduced by Zhou et al.) and colour. The authors employ kernel based learning algorithms (SVM and LS-SVM) to perform the actual classification. The classification scheme works as follows:

- The clinician identifies a relevant frame and marks a circular Region of Interest (ROI) surrounding the suspicious tissue.
- The features are extracted only from the ROI i.e. the pixels outside the circular
ROI are ignored.

- The classification is performed on extracted features, which in turn returns the prediction about the tissue malignancy.

The authors point out that the classification based on the whole images did not produce the successful results. The most obvious remark about this method is that it requires high level of interaction from the clinician, who needs not only to watch the entire video, but also to identify a suspicious frame and then a suspicious region of interest within that frame.

In (Huang et al., 2004), the authors propose a computer-assisted system aimed at detecting *Helicobacter pylori* infection and related inflammations in endoscopy images. In order to make an automated diagnosis, three images from different parts of the stomach (antrum, body and cardia) are acquired, from which the small sub-images of plain tissue are segmented. This is followed by colour and texture feature extraction. The colour features include maximum, average end extension (distribution extent) values of Red, Green, Blue and Gray-scale histograms ($3 \times 4 = 12$ features in total). The texture features are extracted using the spatial gray level co-occurrence matrix (SGLCM) and include contrast, entropy and energy of each of the four previously mentioned channels. These features were associated with the histological parameters (AIS - acute inflammation score, CIS - chronic inflammation score, AT - atrophy, IM - intestinal metaplasia and LF - lymphoid follicles) which were calculated from the biopsy samples taken from the same locations as the images being acquired. These patterns were used to train a conjugate gradient back propagation neural network. The authors report a specificity and sensitivity rates of around 80-90% for estimating values of previously mentioned histological parameters above certain predefined alarm thresholds. Consequently, they conclude that their method can offer comprehensive information about the stomach during or just after the completion of the endoscopy examination and may help to overcome the disadvantages for histology of the localised biopsy.
CHAPTER 3. LITERATURE REVIEW

In (Dhandra and Hegadi, 2005), a scheme for the detection of abnormal endoscopic tissue using a morphological watershed segmentation is proposed. According to the authors’ claim, an abnormal image usually contains a higher number of distinctive regions. These are extracted by segmentation in their method. Having performed the segmentation, an image is considered to be abnormal if it contains more regions than the predefined threshold (set to 5). The authors conclude that their work will require further evaluation and that a classification scheme based on ‘segmented region’ features such as shape, size etc. might be considered in the future.

Specular highlights, which are bright spots of light that appear on the surface of illuminated shiny objects are a very undesirable factor in traditional endoscopy. In (Bochko and Miyake, 2006), the authors propose a method of removing them. After Principal Component Analysis, a Gaussian Mixture Model is used to cluster the data into body reflection and non-body reflection classes. This is followed by the second clustering step, where the non-body reflection cluster is separated into smaller components. Resulting clusters having a maximum norm of the mean vector are considered highlights and consequently the image pixels are labelled. These pixels are assigned new values, obtained using the mapping onto the first eigenvector of the body reflection cluster, by using the first eigenvector of the highlight reflection cluster. The computational time of the method for the image size $81 \times 81$ and 61 spectral components was reported to be around 12s.

3.3 Possibilities of image retrieval in a database of endoscopic images

In (Cauvin et al., 1998), authors describe an attempt to build a database of endoscopic images. The aim of this work was to build a reference database of endoscopic images and video sequences to organize the indexing system for computer-assisted retrieval of
specific images. The database described in this work consisted only of standard endoscopy images. However, a similar approach could be taken when building a database of capsule videos obtained in clinical practice. The WCE image archiving system with communication capabilities has many potential benefits. In (Pilling, 2003), J. R. Pilling describes the benefits of a similar system implemented in Norfolk and Norwich University Hospital which involved radiology images. The author lists benefits such as: data available in many places simultaneously, improved specialist efficiency, faster reporting, instant image availability, comparison with previous results, improved medical staff efficiency and reports available with images.

3.4 Conclusions

In this chapter, we have described a number of Computer Vision algorithms that have been applied to the analysis of WCE videos and presented in the literature. Moreover, some relevant papers regarding conventional endoscopy and colonoscopy has also been reviewed.

As to a comparison of methods applied to WCE and conventional endoscopy and colonoscopy, there is one important remark which must be made. WCE takes place in a completely different clinical scenario to the conventional methods. The latter gives full control of the procedure to the clinician who watches and reports on the video display while the procedure is undertaken. On the occasions when a procedure is recorded, it is rarely for the purpose of further analysis. This can be put in contrast to WCE in which the clinician has no control over the device, cannot view in real time, and as a result of this has to watch the video later, which takes up to two hours of their time. This striking difference has far reaching implications on the use of computer-assisted image analysis methods. Since the conventional procedure is under full clinician’s supervision, we might allow high level of interaction between him/her and the system. For
example in the first method described in Section 3.2.2, the clinician, having identified
the suspicious frame, needs to draw a circular region of interest around the suspicious
tissue in order to allow the system to classify it. Although, this seems to be demanding
a lot from the clinician, it is actually not so time consuming if we consider that
the clinician watches the entire video in real time anyway. Hence, drawing a region
of interest around the suspicious area and classifying it is expedient. The situation is
different however, if we consider the WCE scenario. Here, we want not only to make
the diagnosis more accurate, but also to reduce the time of the entire procedure. This is
difficult using conventional real time methods. As we cannot rely on clinician’s inter-
active feedback to identify suspicious regions, instead we must analyse the entire video
off-line, possibly reducing the number of video frames to be analysed by focusing on
only suspicious regions. It is clear that in these two scenarios the tasks of the clinician
and the intelligent computer system have been exchanged at least to some extent i.e. in
conventional methods the clinician was searching for the suspicious region and asking
the intelligent system for the second opinion and in WCE, the system is pointing out the
suspicious regions to the clinicians leaving the task of the final diagnosis to them. This
of course does not prevent us from using the conventional endoscopy scenario within
WCE i.e. the clinician selects the particular region of interest on which he wants to
have the second opinion (apart from his own) of the computer system. From this, it is
apparent that WCE opened a wide range of new possibilities for computer-assisted im-
age analysis and indeed it can be considered an ideal subject of interest for the medical
image analysis community.

With regard to the WCE image analysis methods described in this chapter, our con-
clusions here must be limited as they only form a part of the body of research, which
will be explored in the following chapters. Nevertheless, a picture that emerges from
this description is very positive for image analysis research. Although there has been
a significant increase in the number of papers published, there is still plenty of scope
for research in this area. The next two chapters will complete the review by looking in detail at the areas of topographic video segmentation and bleeding detection. Moreover, in Chapter 6, we will draw final conclusions about the state-of-art of WCE image analysis and discuss the possibilities of future research directions.
Chapter 4

Topographic video segmentation

As was mentioned in the previous chapter, topographic segmentation has received the largest amount of attention from the computer vision and image processing community. This thesis offers a major contribution in this area.

In Section 4.1, we describe a motivation behind this work. Section 4.2 defines the GI regions and their boundaries with respect to the video segmentation task. The following three sections describe in detail the structure of the video segmentation system (see Figure 4.1). This consists of three major tasks: feature selection/extraction (Section 4.4), single image classification (Section 4.5) and video segmentation (Section 4.6). In Section 4.9, we describe a prototype of the topographic video segmentation system that was installed for evaluation in the Gastroenterology Unit of Norfolk and Norwich University Hospital and in Section 4.10, we summarise the findings of this chapter.

4.1 Motivation

The motivation for this work can be explained by three major factors. The first factor is driven by the requirements of the Given Imaging software - Rapid Reader, which relies on clinicians annotating the first images of the stomach, intestine and colon regions in
Figure 4.1: The structure of the topographic video segmentation methods
order to enable two other features of this software: Suspected Blood Indicator (SBI) and Localization (Fisher et al., 2004) functions. SBI is designed to report the location in the video of areas of active bleeding and is discussed in more detail in Chapter 5. The task of the Localization function is to display the route and relative position of the capsule on a graphical torso model. The second factor is of a more general nature and can be explained by the fact that providing these two locations automatically to clinicians eases their work since it directs them to the most relevant part of the video, in this case the small intestine. The small intestine is the most important region for capsule endoscopy, as it is almost impossible to examine using other methods. Generally, the patient has already been examined in the stomach, upper small intestine and colon, using conventional endoscopy and colonoscopy, but most of the small intestine has yet to be examined. The third factor is that segmenting stomach from the intestine region provides the clinician with so called Gastric and Intestinal Transit Times (GTT and ITT) (see Figure 4.2), which are useful diagnostic cues for the clinician as unusual transit times may signify an underlying pathology. These three factors contribute to the most important advantages that automatic WCE video segmentation provides: first, a significant reduction in the amount of time taken by a clinician to assess a WCE video and hence a reduction to the cost of the WCE procedure; and second, assisting the clinician with the diagnosis.

![Figure 4.2: Gastric and Intestinal Transit Times](image)
CHAPTER 4. TOPOGRAPHIC VIDEO SEGMENTATION

4.2 Video regions and their boundaries

The Gastrointestinal tract (see Figure 2.13) consists of the mouth, oesophagus, stomach, small intestine and large intestine (colon). Once awaked, the capsule endoscope is usually outside the body of the patient for no more than a few seconds, before it is placed in the mouth, where it usually stays for a few seconds, before it is swallowed. After which, it is pushed by peristalsis (peristalsis is the natural muscular motion that pushes food through the GI tract) along the oesophagus, in which it stays for around two seconds. Hence, usually we can see only around 3-5 images taken inside the oesophagus (the capsule transmits two images per second), before the capsule reaches the esogastric junction (EJ) separating the oesophagus from the stomach. Typically, the time spent in the stomach (GTT) is around 15 minutes (for our video dataset $15.8 \pm 13.7$ min.). However, this time can be significantly longer, and it is possible for a capsule to stay in the stomach for several hours before it passes through the pylorus (the valve between the stomach and the intestine). The small intestine is the longest part of the GI tract, and the capsule usually spends around 4 hours in transit through this region (for our dataset $233 \pm 60$ min.). It ends with the ileocaecal valve (IV), which marks the beginning of the colon.

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. Annotating the place in the video where the capsule enters the IV is also difficult, since intestine and colon tissues are very similar and are often contaminated with faecal material that occludes the camera view. Annotating the esogastric junction is relatively the easiest of these three tasks, as the physical features inside the mouth, oesophagus and stomach are visually quite unique and, thus significantly different. Considering these observations, we might predict that any WCE topographic segmentation method should produce accurate results for EJ location, followed by the pylorus and we expect the IV location to be the least accurate.
For the purpose of the video segmentation task, we divided the GI tract into four regions. These regions correspond to the anatomy described above apart from one detail located at the beginning of each video. Since the camera starts acquiring images before it is swallowed, the video contains the images of the outside world, which will have to be classified by any video segmentation algorithm. Hence, the first region of the video is labelled with the name \textit{Entrance} and includes the images taken before the capsule was swallowed and those captured in the mouth and the oesophagus. The following three regions, as described in the previous paragraph, correspond to the anatomical regions known as \textit{Stomach}, \textit{Small Intestine} and \textit{Colon}.

\section{Related work}

When the research described in this thesis was starting in 2003, there were neither publications about topographic video segmentation nor publications regarding capsule endoscopy in computer vision and image processing literature. Since that time, in addition to the research carried out by the authors at UEA and described in this chapter, another research group has started work in the area of capsule endoscopy topographic video segmentation. Their findings have since been published and are described below.

Coimbra et al. (Coimbra et al., 2005) present an attempt to segment a WCE video into meaningful parts. They divide the video into four zones: Entrance ($Z_1$) - consisting of image frames acquired from the mouth and oesophagus as well as those acquired before the capsule is swallowed; Stomach ($Z_2$) - whose limits are determined by the oesogastric junction and the pylorus; small intestine ($Z_3$) - delimited by the pylorus and IV; Colon ($Z_4$) - from IV to the end of the footage. MPEG-7 descriptors (Scalable Colour and Homogeneous Texture) are used as low-level image features (Chang et al., 2001). The classification is performed using a Bayesian classifier, which assigns a topographic location label to each frame in the video. Iteration is used to minimise the
segmentation error, resulting in three parameters that show the positions of transitions between the four previously defined zones. In (Coimbra, Campos and Cunha, 2006b), the authors show that using a Support Vector Classifier instead of the Bayesian approach significantly improves the results, which can be used to estimate the capsule Gastric and Intestinal Transit Times. The authors have later shown that combining content with context features may give an additional boost to WCE video segmentation (Coimbra, Kustra, Cunha and Campos, 2006). Contextual features may include spatial location of the capsule inside the body of the patient and capsule displacement velocity. The authors conclude that such an approach mimics more closely additional expert knowledge that the clinician draws on in order to perform the annotation more accurately.

### 4.4 Feature extraction

In this section we describe the different feature extraction methods which were employed for the WCE video segmentation task. These methods utilise Colour, Texture and Motion information contained in the capsule exams and are described in detail in Sections 4.4.1, 4.4.2 and 4.4.3. A development in the extraction of colour and texture features from the pre-segmented parts of the images that contain only plain (clean, not occluded) tissue is described in Section 4.4.4, and Section 4.4.5 discusses feature compression.

#### 4.4.1 Colour

The distribution of colours in an image provides a useful cue for image indexing and object recognition. The colour distribution histogram is the most commonly used method of representing image colour information (Swain and Ballard, 1991). It is relatively invariant to image scale changes, translation and rotation about the viewing axis, and partial occlusion. WCE videos might be segmented using this method since colour
CHAPTER 4. TOPOGRAPHIC VIDEO SEGMENTATION

information is the primary feature analysed by the clinician.

Visually, the mouth contains unsaturated colours, the stomach pinkish colours; the small intestine pinkish to yellowish colours; and the colon pinkish to yellowish colours generally occluded by varying amounts of yellowish to greenish colours caused by faecal contamination.

WCE videos consist of images stored as $RGB$ triplets. In (Berens, 2002), it was shown that $RGB$ colour space is not the best choice for image classification and therefore the use of other perceptually relevant colour spaces was suggested. One of these spaces - $HSI$ (Hue, Saturation, Intensity) (Gonzalez et al., 2004) was shown to provide excellent classification results. Consequently, we use $HSI$ colour space in this work as well. There is a great deal of intensity variation in WCE images as the distance between the WCE and the intestine surface constantly varies. Therefore, we form 2-D $HS$ histograms by ignoring intensity information, forcing intensity invariance, and also reducing data size. The range of colours present in WCE images is relatively small, mapping to a region covering just around 20% of the possible $HS$ colour space, and so we equalise the histograms to this subset (of red to yellowish-green) colours. Figure 4.3 shows typical WCE images acquired from the mouth, stomach, intestine and colon regions, and their respective $HS$ histograms. It can be seen that the colour distribution of the stomach is slightly shifted towards red, compared to the distribution within the intestine. It is also clear that the colour distribution of the colon tissue is highly similar to that of the small intestine, when it is free of faecal contamination. However, colon images are generally obscured by the presence of faecal contamination which has a distinct hue-saturation signature.

4.4.2 Texture

A short inspection of capsule images suggests that texture features can play an important role in the topographic video segmentation. The most prominent texture pattern
that distinguishes different organs is the villi (see Figure 4.3 C) - the small finger-like projections, responsible for food absorption, which are present in the small intestine, but not in the neighbouring regions of stomach and colon. Texture patterns in the mouth and oesophagus are also very distinctive. Moreover, there are a number of fluids such as saliva, bile and digestive remains, each possessing a different textural pattern. These differences can be represented using textural features, which are described in this section.

Texture has been often used as a feature in medical image analysis techniques (Howarth et al., 2005; Coimbra et al., 2005; Vilarinao et al., 2005; Wang et al., 2001; Xu et al., 2006). In most medical imaging systems, textural features are calculated from a grey-scale image. WCE is a colour imaging technology and therefore, a more recent method involving colour image texture can be employed. The Local Binary Pattern operator (LBP) (Mäenpää and Pietikäinen, 2004; T. Ojala and Mäenpää, 2002), is one of many colour texture methods that has proven to perform well in texture classification applications. This technique has been implemented, but only in grey-scale images, in
existing image retrieval systems (Mäenpää and Pietikäinen, 2005). In this work, we use a new method of calculating LBP recently introduced by Connah and Finlayson (Connah and Finlayson, 2006), that characterises the images by their 3D LBP histograms. This method is an extension of the work of Mäenpää and Pietikäinen (Mäenpää and Pietikäinen, 2004), who calculated LBP histograms for the three colour channels independently. The results demonstrate that computing a 3D histogram retains more information about the image, thus yielding superior performance.

Let us consider a grey-scale image whose intensity can be written \( I(x, y) \). The \( N \) neighbours of any given pixel \( p \) can be denoted as \( n_i, i = 0, \ldots, N - 1 \). In order to calculate an LBP value, the value of each neighbour \( n_i \) is compared to the value of \( p \) to establish whether it is greater than or less than \( p \). This can be written as a function mapping each \( n_i \) onto a value \( b_i \) as follows:

\[
b_i = \begin{cases} 
1 & \text{if } n_i \geq p \\
0 & \text{if } n_i < p
\end{cases}
\]  

(4.1)

The LBP value for pixel \((x_0, y_0)\) is calculated by concatenating the \( N \) binary values \( b_i \) into an \( N \)-bit number, which can be described as follows:

\[
LBP(x_0, y_0) = \sum_{i=0}^{N-1} b_i 2^i
\]  

(4.2)

The notion of LBP can be extended into a colour image by calculating the LBP value in each of three channels separately, resulting in a triplet of LBP values for each pixel \( LBP_k(x, y) \) where \( k \in \{R, G, B\} \). After calculating an LBP value at every pixel for each channel, the image can be characterised by the histogram built from LBP values. This can be built in two ways: the first method involves concatenating the three separate histograms (Mäenpää and Pietikäinen, 2004), and the second is derived from the 3D joint histogram, which preserves more information about the interactions between the channels (Connah and Finlayson, 2006), thus we expect it to be superior to the first
method.

As in (T. Ojala and Mäenpää, 2002), we use a circular neighbourhood around the central pixel \( p \), sampling at points which are equidistant from \( p \) and one another (see Figure 4.4). When a sample point does not fall precisely on a pixel, its value is calculated using bilinear interpolation.

Moreover, following (T. Ojala and Mäenpää, 2002), we calculate \( LBP_{ri} \), which is invariant to image rotations and reduces the number of possible patterns:

\[
LBP_{ri} = \min \{ ROR(LBP, i) | i = 0, 1, ..., N - 1 \} \quad (4.3)
\]

where the \( ROR \) function shifts the N-bit binary LBP value, \( i \) bits to the right, with wrap-around.

The number of possible patterns can be further reduced by considering only patterns of a specific type. This is achieved by introducing an extra constraint on pattern uniformity (T. Ojala and Mäenpää, 2002). Given the binary string \( b_0b_1...b_{N-1} \), the pattern uniformity is defined as the number of transitions that occur in that string and each transition a change from 0 to 1 or vice-versa including wrap-around.
\[ \text{Uniformity} = ||b_0 - b_{N-1}|| + \sum_{i=0}^{N-2} ||b_i - b_{i+1}|| \] (4.4)

Following (T. Ojala and Mäenpää, 2002; Connah and Finlayson, 2006), we consider only patterns with \( 0 < \text{uniformity} \leq 2 \), as these patterns are relevant and approximately correspond to edges, line endings and corners (as in Figure 4.4). In the case of 8 sampling points, there are 7 unique patterns, \( 21(7 \times 3) \) bins for the independent histogram and \( 343(7^3) \) for the joint histogram.

We then apply Principal Component Analysis, which reduces the feature vector length.

### 4.4.3 Motion

The work described in Section 3.1.3 suggests it is possible to use motion features for discriminating between different gastrointestinal zones. A brief observation of any capsule video clearly suggests that the motion patterns differ in different organs. At the entrance, we can distinguish two phases: the first phase takes place when the capsule is outside the body of the patient, in which case we cannot predict much about its motion since the capsule may be waiting to be picked up and swallowed or in the hand of the clinician or the patient in which case, there would be some irregular motion pattern. The second phase begins once the capsule is swallowed by the patient and the capsule travels through the mouth and down the oesophagus with a distinct and quick movement. In the stomach, the capsule first tumbles around, then reaches the bottom of the stomach, where it moves according to the contractions of the pylorus until it traverses the valve and enters the small intestine. In this the longest GI organ, the capsule moves within a much more confined, tube-shaped space pushed by the regular intestinal contractions. Having passed the IV, the capsule enters the colon, where it usually drops to the bottom of the ascending colon, after which it usually exhibits a much more slow transit speed.
than in the small intestine.

In this section, we describe a feature extraction method which explores these differences in motion patterns, in order to allow us to distinguish between different GI regions.

Our method of motion feature extraction comprises of three steps. In the first step, we extract motion vectors of 64 \(16 \times 16\) pixel image sub-blocks as shown in Figure 4.5. There are a number of block matching algorithms, which can be employed to carry out this task (Lu and Liou, 1997; Po and Ma, 1996; Zhu and Ma, 2000; Nie and Ma, 2002). The idea behind all these methods is to divide the current frame into a grid of macro blocks that are then compared with the corresponding blocks and its adjacent neighbours in the previous frame in order to create a vector that would describe the movement of the macro block from a location in the current frame to the location in the previous frame. The search area is constrained by the search parameter \(p\) so that only \(p\) pixel displacements of the corresponding macro block in the previous frame are allowed as can be seen in Figure 4.6. The larger the parameter \(p\), the more computational expensive the motion estimation process is. The candidate macro block is chosen as the best match if the output of the cost function is minimal. There are various cost functions which include Mean Absolute Difference (MAD, used in this work and given by Equation 4.5), Mean Squared Error (MSE) or Peak-Signal-to-Noise-Ratio (PSNR).

\[
MAD = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |C_{ij} - P_{ij}|
\]  

(4.5)

where \(N\) is the size of a macro block and \(C_{ij}\) and \(P_{ij}\) are the corresponding pixels being compared from a current macro block and a macro block from the previous frame.

In this work we calculate the motion vectors using the recent Adaptive Rood Pattern Search (ARPS) (Nie and Ma, 2002). For details of ARPS, the readers are advised to refer to (Nie and Ma, 2002). We used \(p = 7\) as the search parameter and MAD as the cost function.
Figure 4.5: A grid of one hundred $16 \times 16$ macro blocks used for motion feature extraction.

Figure 4.6: A macro block of size sixteen and a search parameter of size seven.

Figure 4.5 shows the grid of macro blocks plotted onto a sample WCE image that is used for motion feature extraction. In Figure 4.7 we can see a sample grid of motion vectors, which was calculated from the consecutive images of one of the WCE videos. We can see that in this particular example the motion was generally from the left side of the image to the right.

From the set of motion vectors, denoted as $x$, we calculate the following three features:
CHAPTER 4. TOPOGRAPHIC VIDEO SEGMENTATION

1. sum of motion vectors

\[ F_1 = \sum_{i=1}^{N} x_i \]  \hspace{1cm} (4.6)

2. sum of vector lengths

\[ F_2 = \sum_{i=1}^{N} |x_i| \]  \hspace{1cm} (4.7)

3. sum of the following dot products

\[ F_3 = \sum_{i=1}^{N} (x_i \cdot u_i) \]  \hspace{1cm} (4.8)

where \( u_i \) is a unit length vector, whose direction depends on its location on the grid shown in Figure 4.8 and \( N \) denotes the number of macro blocks. Hence, the value of \( F_3 \) describes the motion of the capsule in terms of whether it is moving forward or backward. A positive \( F_3 \) will suggest that the capsule is moving forward and the negative will suggest the opposite.
CHAPTER 4. TOPOGRAPHIC VIDEO SEGMENTATION

Analogically to $F_1$, $F_2$, $F_3$, we also calculate three more features denoted as $F_4$, $F_5$, $F_6$ representing standard deviations of the respective expressions.

Motion information extracted from two consecutive images of the WCE video and contained in features $F_1 - F_6$ cannot tell us much about their location with respect to different GI organs as the motion patterns are very chaotic and the particular set of feature values can appear in any of the digestive organs. Therefore, we combine motion features over a larger set of 41 consecutive frames and analyse the entire sequence. We carry this out by forming vectors $s_1 - s_6$ where $s_n = (F_n^i - 20, F_n^i - 19, ..., F_n^i, ..., F_n^i + 19, F_n^i + 20)$ and $i$ denotes the number of the frame in the video sequence for which the feature vectors $s_n$ are formed.

Each of the $s_1 - s_6$ feature vectors is transformed using the Discrete Fourier Transform (DFT) (DeFatta et al., 1988), which is given by Equation 4.9.

$$X(k) = \frac{1}{N} \sum_{i=0}^{N-1} x(n) e^{-j\pi nk/N} \quad k = 0, 1, ..., N - 1$$ (4.9)
Vectors $s_1 - s_6$ after transformation consist of real and imaginary parts. In our application, we will consider only their magnitudes and denote them $S_1 - S_6$.

The DFT has a particular property which is very useful for our application and is known as the circular shift of the DFT input (DeFatta et al., 1988):

$$DFT[x(n + m)] = X(k)e^{j2\pi km/N}$$ \hspace{1cm} (4.10)

As stated previously, we consider only the magnitudes of the DFT transform. Since the second term of Equation 4.10 has a unit magnitude, we can write:

$$|DFT[x(n + m)]| = |X(k)|$$ \hspace{1cm} (4.11)

From the above equation, we can see that the circular shift of the input does not affect the magnitude of the output coefficients. This is an important property that should allow us to distinguish between different motion patterns, which in our case are not aligned.

Another DFT property exploited in our application uses the fact that our inputs are real numbers. This forces the real part of the DFT output to be symmetric and the imaginary part antisymmetric (DeFatta et al., 1988). Hence, we can infer that our DFT magnitudes are symmetric and thus discard almost a half of transform coefficients (in our case 41 samples, with 21 output coefficients retained and 20 discarded).

### 4.4.4 Sub-image region (SubIR) selection

WCE images are often obscured (to a varying degree) by strong shadows, or by air-bubbles and other artifacts - such as mucus, bile, faeces, food etc. Histograms built using the entire image will contain any visual contamination present in the image. To address this problem, we want to extract only those parts of the image which contain only plain tissue. In order to do this, we divide each WCE (256 × 256 pixels) image...
into a grid of 28 sub-images, $32 \times 32$ pixels each, covering most of the image area, as shown in Figure 4.9.

We derive five parameters for each of the sub-images - Mean Intensity, Saturation, Hue, and Standard Deviation of Intensity and Hue. The values for these parameters were set manually following an experiment involving an observer (the author of this thesis) adjusting thresholds so that sub-images containing visual contamination (i.e. outside the expected colour range for the tissue type) are rejected. The following set of parameter thresholds was selected ($0 < m_H < 0.08; 0.2 < m_I < 0.7; 0.25 < m_S < 0.95; \sigma_H < 0.005; \sigma_I < 0.075$) by the observer. We tested each sub-image against the five parameters and discard those that fall outside the range of values typical for visually clear images of gastrointestinal tissue. The remaining sub-images form a sub-image region (SubIR) that will be used in a feature extraction process. Figure 4.10 shows eight typical images acquired in the stomach and intestine showing only those sub-images selected by the algorithm described above. Consequently, the features described in Sections 4.4.1 and 4.4.2 are extracted from SubIR segmented in the manner described above. Table 4.1 shows the mean and standard deviation of the number of images extracted from the data set with at least one SubIR. It can be seen that the
largest number of images containing SubIR comes from the stomach and intestine and therefore when SubIR features are used, we would expect the most significant impact on classification of these two video regions. The small number of images in both entrance and colon can be respectively explained by the relatively long time intervals before the capsule is swallowed and the occlusion by digestion remains in the last part of the GI track. As the results in Sections 4.7.1 and 4.8.1 will show, combining colour and texture information from the SubIR with the entire image improves video segmentation results, particularly in case of stomach/intestine discrimination.

**Figure 4.10:** WCE images showing selected SubIRs. A-D Stomach; E-H Intestine.

**Table 4.1:** The mean and standard deviation of the number of images with at least one SubIR for different GI parts.

<table>
<thead>
<tr>
<th></th>
<th>entrance</th>
<th>stomach</th>
<th>intestine</th>
<th>colon</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>0.12</td>
<td>0.89</td>
<td>0.74</td>
<td>0.19</td>
</tr>
<tr>
<td>σ</td>
<td>0.22</td>
<td>0.14</td>
<td>0.13</td>
<td>0.22</td>
</tr>
</tbody>
</table>
4.4.5 Compression

The Hybrid Transform

The 2-D histogram feature vector is a large structure, and therefore for practical applications, compression is necessary. Hybrid transforms are widely used in image compression since they have potential for greater compression than single coding techniques. An example of such a method is the JPEG compression where the Discrete Cosine Transform (DCT) is combined with entropy coding. In this work we compress the $HS$ histograms using a hybrid transform consisting of DCT compression followed by Principal Component Analysis (PCA) (Berens, 2002; Berens and Finlayson, 2000). The reason for using the hybrid transform instead of PCA alone follows from the fact that our 2D histograms are large structures and it would require an impractically large number of histograms to provide a reliable basis decomposition. To address this problem, the frequency coding (DCT) employed in the first-stage of the hybrid transform significantly reduces the size of the data, which can be further compressed in the second stage using the optimal transform (PCA). The hybrid transform produces a small feature vector that provides a fast and accurate model of the histogram data and was found to perform better in colour indexing experiments than PCA alone (Berens, 2002). In fact, it was shown that for a large colour image database (MPEG-7, 5466 images), it was enough to use as few as 8 principal components as the feature vector to obtain excellent image retrieval results. Figure 4.11 shows the first three principal components calculated using the Hybrid Transform (DCT followed by PCA) from 1000 $HS$ histograms extracted from one WCE video. Each dot on the graph represents one histogram. It can be seen that in this video there is a distinct separation between the colour distributions of the stomach and the intestine, whereas one cannot actually see the separation between the intestine and colon.

In this work, PCA alone is also used to compress 3-D LBP histograms and DFT
Figure 4.11: Three first principal components representing compressed histograms extracted from four different video regions.

In the two following subsections, we give the mathematical details of DCT and PCA.

**Discrete Cosine Transform (DCT)**

The DCT (Gonzalez et al., 2004) is one of many frequency transforms. It represents a signal (can be 1-D or higher) as a linear sum of frequency cosine basis functions. It has been used in a number of applications requiring compression including JPEG image compression, MJPEG, MPEG, and DV video compression. Moreover, under reasonable statistical assumptions, the DCT has been shown to approach the Karhunen-Loève transform (PCA). Below, the 1-D DCT is given.

\[
X_k = \alpha(k) \sum_{n=0}^{N-1} x_n \cos \left[ \frac{\pi}{N} \left( n + \frac{1}{2} \right) k \right]
\]  

(4.12)
where \( k = 0, \ldots, N - 1 \) and

\[
\alpha(k) = \begin{cases} 
\frac{1}{\sqrt{N}} & k = 0 \\
\sqrt{\frac{2}{N}} & 1 \leq k \leq N - 1 
\end{cases} \tag{4.13}
\]

Multidimensional versions of DCT follow from the one-dimensional definition since they are a composition of DCTs along each dimension e.g. a 2-D DCT of an image is the 1-D DCT, as above, performed along the rows and then along the columns (or vice versa). Assuming the square image \( N \times N \), the normalised version of the 2-D DCT can be written as follows:

\[
X_{u,v} = \alpha(u)\alpha(v)\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} x_{m,n} \cos \left( \frac{2n + 1}{2N} u\pi \right) \cos \left( \frac{2m + 1}{2N} v\pi \right) \tag{4.14}
\]

where \( u = 0, \ldots, N - 1; \quad v = 0, \ldots, N - 1 \) and \( \alpha(u), \alpha(v) \) are calculated as in Equation 4.13.

In Chapter 5, we will also use the 3-D DCT, which can be constructed by a straightforward extension of the 2-D version:

\[
X_{u,v,w} = \alpha(u)\alpha(v)\alpha(w)\sum_{l=0}^{N-1}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} x_{m,n} \cos_{\text{prod}} \tag{4.15}
\]

where

\[
\cos_{\text{prod}} = \cos \left( \frac{2n + 1}{2N} u\pi \right) \cos \left( \frac{2m + 1}{2N} v\pi \right) \cos \left( \frac{2l + 1}{2N} w\pi \right) \tag{4.16}
\]

In case of 2-D DCT, the coefficients are ordered in terms of information importance according to a zigzag pattern as shown in Figure 4.12. Hence, the unwanted frequency coefficients can be removed using a zonal mask acting as a low-pass filter. Here, the
reduced coefficient set is created by retaining in the numerical order the gray coloured coefficients and discarding those masked by zeros.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>6</th>
<th>7</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 4.12:** Compression zonal mask applied to $8 \times 8$ coefficient block. The mask removes coefficients 11 to 64

As to 3-D DCT, the most important coefficients are in the pyramid shaped region in the corner of the transformed cube (see Figure 4.13). Analogically to the 2-D case, this pyramid can act as a zonal mask in a process of unwanted coefficients removal.

![Zonal mask in 3-D](image.png)

**Figure 4.13:** Zonal mask in 3-D.

**Principal Component Analysis**

The PCA (Pratt, 2001; Gonzalez et al., 2004) of a distribution is also called the (discrete) Karhunen-Loève transform (KLT) or the Hotelling transform. It is the only optimal linear transformation for keeping the subspace that has largest variance. However, computational complexity of PCA ($O(M^2N^2)$) is higher when compared, for example to the DCT discussed above ($O(MN \log 2MN)$). The basis functions are called the
principal components (PCs) which is the convention that will be followed in the rest of this thesis. It is worth noting that unlike other linear transforms (including DCT), PCA does not have a fixed set of basis vectors. It computes the mean vector $m_x$ and forms the covariance matrix, $C_x$ for a distribution $x$, of size $K$:

$$m_x = \frac{1}{K} \sum_{k=1}^{K} x_k$$  \hspace{1cm} \text{(4.17)}$$

$$C_x = \frac{1}{K} \sum_{k=1}^{K} (x_k - m_x)(x_k - m_x)^T$$  \hspace{1cm} \text{(4.18)}$$

Having calculated $m_x$ and $C_x$, the PCA of a distribution $x$ is given by:

$$y = A(x - m_x)$$  \hspace{1cm} \text{(4.19)}$$

where the rows of matrix $A$ are the normalised eigenvectors of $C_x$, ordered according to the decreasing corresponding eigenvalues. The covariance matrix of $y$ is a diagonal matrix $C_y$, whose diagonal elements are the eigenvalues of $C_x$. The inverse transformation (since $A$ is orthonormal, its inverse equals its transpose) produces the reconstructed $x$:

$$x = A^T y + m_x$$  \hspace{1cm} \text{(4.20)}$$

With regard to compression, the usefulness of PCA becomes clear when only some subset of $q$ eigenvectors is used, in which case $A$ becomes a $q \times n$ matrix $A_q$. Hence, the approximated reconstruction can be given as follows:

$$\hat{x} = A_q^T y + m_x$$  \hspace{1cm} \text{(4.21)}$$

There is an alternative way of calculating PCA which utilises singular value decomposition (Golub and van Loan, 1983) and was used in this work. Here, $x$ denotes the
distribution with the subtracted mean, where each column contains a different subject, and each row different variable. Then we can write PCA as:

\[ y = U^T x = SV^T \]  \hspace{1cm} (4.22)

where \( USV^T \) is the singular value decomposition of \( x \).

### 4.5 Single image classification

As Figure 4.1 shows, before the video can be segmented into meaningful parts, the set of single images (represented using the extracted features) has to be classified as belonging to one of the four classes. There is a number of classifiers that can be employed for this task (Webb, 2002). In this work, we test three of them: k Nearest Neighbour (kNN), Multivariate Gaussian and Support Vector Classifier (SVC). The former is used in the initial tests since it is arguably the most popular classifier and is easy to implement. On the other hand, Support Vector Classifiers have received a lot of interest in recent years and are used in the state-of-art pattern recognition applications. Hence, it was suspected that SVC may yield better results than kNN and that is why it was chosen for the final experiments. The Multivariate Gaussian classifier is also tested for the reason that it provides the result of the classification as the probability, which is used in the later video segmentation stage as the input of the Viterbi algorithm (see Section 4.6.3). Incidentally, although the standard (binary) SVC classifier does not provide the classification result as class probabilities, we derive them using the pairwise coupling method described in the last part of Section 4.5.3.

The following three subsections, describe in detail kNN, Multivariate Gaussian and SVC classifiers respectively.
4.5.1 K Nearest Neighbour (kNN)

K-Nearest Neighbor (kNN) (Webb, 2002) classification is a simple to implement, yet powerful classification method. The fundamental idea behind kNN classification is that similar observations belong to the same class. Thus, one simply has to count the class labels of a certain number (called \( k \)) of the nearest neighbours to assign a class label that scored the highest count (see Figure 4.14).

\[ \text{Figure 4.14: A figure illustrating the kNN algorithm - a classified vector with the k=5 closest neighbours pointed by arrows.} \]

The number of the nearest neighbours, \( k \), is usually odd in order to avoid ties. However, should the number of neighbours be even, there are ways to break ties e.g. one might assign the label of the class whose members are closer to the classified vector (the shortest mean distance to the nearest neighbours) or of the class that is more compact (the shortest distance to the last of the nearest neighbours).

The optimal choice of \( k \) depends on the data; generally, larger values of \( k \) reduce the effect of noise on the classification. However, they can make the boundaries between
classes less distinct.

It can be shown that the error rate of a kNN classifier is bounded by twice the Bayes error rate. The kNN algorithm is conceptually simple to calculate as each classification requires \( n \) calculations where \( n \) denotes the size of the training set. For large \( n \) however, it may become computationally excessive. There are algorithms that try to reduce the influence of this drawback e.g. by pre-computing the distance matrix. The topic is further explored in (Webb, 2002).

### 4.5.2 Multivariate Gaussian

The next classifier that was used in this work is a Multivariate Gaussian (MG). Given a feature vector \( x \in \mathbb{R}^N \),

\[
x = [x_1, x_2, ..., x_N]
\]  

(4.23)

we can estimate the multivariate Gaussian probability density (assuming the covariance matrix \( \Sigma \) is non-singular)

\[
p(x | \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^N|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}
\]  

(4.24)

where \( \mu \) is a class mean vector, which is obtained by averaging coefficient values \( x \) for all images belonging to that class from the training set; \( \Sigma \) is the covariance matrix and \( | \cdot | \) denotes the determinant.

A Multivariate Gaussian distribution can be thought of as a generalization in higher dimensions of the one-dimensional Gaussian distribution, whose familiar equation follows:

\[
p(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]  

(4.25)
where $\sigma$ denotes the variance.

A MG classifier assigns the class label of the distribution which produced the highest probability.

### 4.5.3 Support Vector Classifier

Support Vector Machines (SVM) (Gunn, 1998; Vapnik, 1995, 1998; Webb, 2002) were proposed by Vapnik (Vapnik, 1995) to address the problem of empirical data modelling which is very common in engineering applications. They are becoming widely used since they have many attractive features and promising empirical performance. The terms SVC and SVR refer to Support Vector Classification and Support Vector Regression respectively. SVM map pattern vectors to a high-dimensional feature space where a maximal margin (best separating) hyperplane is constructed (see Figure 4.15).

Consider a binary classification task in which a set of training patterns $\{x_i, i =
1, ..., n} are assigned to one of two classes \( \omega_1 \) and \( \omega_2 \), with corresponding labels \( y_i = \pm 1 \). Then

\[
g(x) = w^T x_i + w_0
\]  

(4.26)

may be considered a linear discriminant function, where \( w \) denotes a weight vector. Hence, all training points are correctly classified if

\[
\forall y_i (w^T x_i + w_0) > 0
\]

The distance between each of two canonical hyperplanes \( H_1 \) and \( H_2 \) and the separating hyperplane \( A : g(x) = 0 \) is \( 1/|w| \) and is called the margin, thus

\[
\begin{align*}
w^T x_i + w_0 & \geq +1 \quad \text{for} \quad y_i = +1 \\
w^T x_i + w_0 & \leq -1 \quad \text{for} \quad y_i = -1
\end{align*}
\]  

(4.27)

The points that lie on the canonical hyperplanes are called support vectors (see Figure 4.15).

Maximizing the margin means seeking the solution that minimizes \( |w| \) subject to the constraints

\[
y_i (w^T x_i + w_0) \geq 1 \quad i = 1, ..., n
\]  

(4.28)

Lagrange formalism is a standard approach to optimization problems with equality and inequality constraints. The primal form of the objective function follows:

\[
L_p = \frac{1}{2} w^T w - \sum_{i=1}^{n} \alpha_i (y_i (w^T x_i + w_0) - 1)
\]  

(4.29)

where \( \{\alpha_i, i = 1, ..., n; \alpha \geq 0\} \) are the Lagrange multipliers.

The solution to the problem of minimizing \( w^T w \) subject to constraints (4.28) is equivalent to determining the saddlepoint of \( L_p \), at which \( L_p \) is minimized with respect
to \( w \) and \( w_0 \) and maximized with respect to \( \alpha_i \). Differentiating \( L_p \) with respect to \( w \) and \( w_0 \) and equating to zero yields

\[
\sum_{i=1}^{n} \alpha_i y_i = 0
\]

\[
w = \sum_{i=1}^{n} \alpha_i y_i x_i
\]  \hspace{1cm} (4.30)

Substituting into equation (4.29) gives the dual form of the Langrangian

\[
L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j
\]  \hspace{1cm} (4.31)

which is maximized with respect to \( \alpha_i \) subject to

\[
\alpha_i > 0
\]

\[
\sum_{i=1}^{n} \alpha_i y_i = 0
\]

It can be shown that support vectors always have non-zero Lagrange multipliers.

Numerical quadratic programming solvers can be employed to obtain Lagrange multipliers. Having done that, the value of \( w_0 \) may be found from

\[
\alpha_i (y_i (x_i^T w + w_0) - 1) = 0
\]

using any of the support vectors. The solution for \( w \) is given by (4.30):

\[
w = \sum_{i=1}^{n} \alpha_i y_i x_i
\]

A new pattern \( x \) will be classified according to the sign of
In case of linearly non-separable data (see Figure 4.16), we must relax the constraints (4.28) by introducing 'slack' variables $\xi_i, i = 1, ..., n$

$$w^T x_i + w_0 \geq 1 - \xi_i \quad for \quad y_i = +1$$

$$w^T x_i + w_0 \leq -1 + \xi_i \quad for \quad y_i = -1$$

$$\xi_i \geq 0 \quad i = 1, ..., n \quad (4.32)$$

We replace $w^T w/2$ by $w^T w/2 + C \sum \xi_i$ where $C$ is a 'regularization' parameter - the lower the value of $C$ the smaller the penalty for 'outliers' and 'softer' margin. It can be shown that minimizing $w^T w/2 + C \sum \xi_i$ does not affect the dual form of the Lagrangian
which is the same as linearly separable data classifier (4.31). $L_D$ is maximized with respect to $\alpha_i$ subject to

$$0 \leq \alpha_i \leq C$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Hence, the only change to the maximisation problem is the upper bound of $\alpha_i$.

It is common to apply a support vector algorithm in a transformed feature space $\phi(x)$, where $\phi$ denotes a non-linear function. In that case the linear discriminant function (4.26) becomes

$$g(x) = w^T \phi(x) + w_0$$  \hspace{1cm} (4.34)

The dual form of the Lagrangian (4.31) becomes

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j)$$ \hspace{1cm} (4.35)

Scalar products between feature vectors can be replaced by a kernel function.

$$K(x, y) = \phi(x)^T \phi(y)$$ \hspace{1cm} (4.36)

A set of possible kernels is presented in Table 4.2.

**Conditional probability estimation**

The disadvantage of SVCs is the fact that they do not provide the posterior class probabilities that are needed by our application in the video segmentation stage as the input
Table 4.2: Support vector machine kernels

<table>
<thead>
<tr>
<th>Nonlinearity</th>
<th>Mathematical form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polynomial</td>
<td>((\gamma x^T y + r)^d)</td>
</tr>
<tr>
<td>Gaussian</td>
<td>(\exp(-\gamma</td>
</tr>
<tr>
<td>Exponential</td>
<td>(\exp(-\gamma</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>(\tanh(\gamma x^T y + r))</td>
</tr>
</tbody>
</table>

of the Hidden Markov Model (see Section 4.6.3).

Given training data \(x_i \in \mathbb{R}^N\), labelled by \(y_i = \{-1, 1\}\), the binary SVC calculates a decision function \(f(x)\) so that \(\text{sign}(f(x))\) is the prediction of any test data.

There are a number of approaches approximating these probabilities (Platt, 1999; Lin et al., 2003; Rüping, 2004). Platt in (Platt, 1999) proposes to approximate \(p(y = 1|x)\) by a following sigmoid function where \(f(x)\) denotes the SVC decision function.

\[
P(y = 1|x) = \frac{1}{1 + \exp(Af(x) + B)}
\] (4.37)

Our application uses the algorithm proposed by Lin et al. (Lin et al., 2003), which is an improved version of Platt’s original algorithm. In order to avoid bias for the training data, \(A\) and \(B\) are estimated on a cross-validation set. To estimate values of \(A\) and \(B\), we choose subset of \(l\) training data \((N_+\) of them with \(y_i = 1\) and \(N_-\) of them with \(y_i = -1\)) and solve the following maximum likelihood problem:

\[
\min_{z = (A, B)} F(z)
\] (4.38)

where

\[
F(z) = -\sum_{i=1}^{l} (t_i \log (p_i) + (1 - t_i) \log (1 - p_i))
\] (4.39)
\[ p_i = \frac{1}{1 + \exp(\alpha f(x_i) + B)} \quad \text{and} \quad t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2} & \text{if} \quad y_i = 1 \\ \frac{1}{N_- + 2} & \text{if} \quad y_i = -1 \end{cases}, \quad i = 1, 2, ..., l \]

The number of classes in our problem is four, thus the use of multi-class classifier is necessary. Our approach uses the widely used multi-class classifier design called *

pairwise coupling* (Refregier and Vallet, 1991; Price et al., 1995; Hastie and Tibshirani, 1998; Wu et al., 2004) which combines together all pairwise comparisons for each pair of classes obtained using binary SVC classifiers. In our application we use a method described in (Wu et al., 2004), which can be considered as an extension of (Refregier and Vallet, 1991). In this method we solve the following optimisation formulation:

\[
\min_p = \frac{1}{2} \sum_{i=1}^{k} \sum_{j \neq i} (r_{ij} p_i - r_{ij} p_j)^2 
\] (4.40)

subject to

\[
\sum_{i=1}^{k} p_i = 1, p_i \geq 0, \forall i
\] (4.41)

where \( r_{ij} \) are the estimated pairwise class probabilities and \( p_i \) the probabilities that we want to estimate.

Once we have the posteriors, we can replace the Gaussians in the HMM model with the SVCs.

### 4.6 Video segmentation

The final stage of the topographic video segmentation involves performing the actual segmentation on the sequence of single image classification results (see Figure 4.1). There are a number of algorithms that can be used to perform this task. In this work,
three such methods have been investigated. The first one, described in Section 4.6.1 is the simplest and naïve. It was tested in this work only for comparison purposes with the two other more accurate methods which utilise the Sliding Window and Viterbi algorithm respectively.

The rest of this section describes all three methods in more detail.

### 4.6.1 A naïve segmentation algorithm based on converging search

This fast algorithm (see Figure 4.17) searches the boundary between two GI regions. It starts from the beginning of the video with the entrance/stomach classifier. Having classified the first frame, it moves forward or backward depending on the outcome of the classification, gradually decreasing the step and in case of no misclassifications it converges at the esogastric junction (the valve between the oesophagus and stomach). Then, a random frame behind and in the vicinity of the EJ is chosen and the procedure is repeated with the stomach/intestine classifier, resulting in obtaining the pylorus position. And finally, the procedure is repeated again with the intestine/colon classifier, which returns the position of the IV. This algorithm, however fast, is prone to large segmentation errors which are the results of the misclassification in the early iteration.

It is possible to improve the results of the classification using this method by e.g. running the algorithm several times for different initial and/or different steps and then using the arithmetic mean or the median of the several candidates to calculate the final position of the transition point.

### 4.6.2 Sliding Window

The sliding window method has been widely used in video segmentation, particularly in shot detection applications. It is designed to provide better results than a naïve segmentation algorithm at the expense of longer video processing time. Figure 4.18 illustrates
how the sliding window algorithm can be used for WCE topographic video segmentation. First the entrance/stomach classifier is used to detect the pylorus position. This is followed by the stomach/intestine and finally intestine/colon discrimination. As the results in Section 4.8.1 will show, there is some improvement in the sliding window segmentation compared to the simple naïve method.

Formally, we define the result sequence \( S = \{S(m) : S(m) \in \{-1, 1\} \land m \in \mathbb{N} \land m \leq M\} \) where \( M \) denotes a length of the classification sequence (sliding window). From the result sequence \( S \), we calculate the estimate classification error \( E \) as a sequence:

\[
E(i) = \sum_{k=1}^{i} S_{pos}(k) - \sum_{k=i+1}^{M} S_{neg}(k)
\]

where

\[
S_{pos}(k) = \begin{cases} 
1 & S(k) = 1 \\
0 & S(k) = -1
\end{cases}
\]
Figure 4.18: Sliding window video segmentation. E,S,I and C denote entrance, stomach, intestine and colon respectively.

and

\[ S_{\text{neg}}(k) = \begin{cases} -1 & S(k) = -1 \\ 0 & S(k) = 1 \end{cases} \]

The minimum of sequence \( E \) determines the most probable position of the transition point, as seen in Figure 4.18. If the transition point is in the second half of the sequence, the sliding window is shifted to the right and the error sequence is recalculated. When the transition point is found (\( \min(E) \) in the first half of the sequence), the algorithm switches to the next classifier and repeats the operation until it finds all the transition points or reaches the end of the footage.

4.6.3 Hidden Markov Model

Hidden Markov Models (HMM) (Levinson et al., 1983; Rabiner, 1989) are widely used in video segmentation. In (Boreczky and Wilcox, 1998), HMM are used to segment the
video into separate shots. The transitions between shots are modelled using different states: fade, dissolve, cut, pan or zoom. In (Xie et al., 2003), the authors present a system that segments a soccer video into parts, representing active playing sequences and breaks.

Figure 4.19 shows the HMM we designed for capsule endoscopy video segmentation. We use a simple left-to-right model (Rabiner, 1989) consisting of four states: Entrance, Stomach, Intestine, Colon. Each arc in Figure 4.19 has an associated probability of either a transition to the next state or remaining in the same state. The probability distribution that models the WCE image features is associated to each state. The input of the HMM is the sequence of conditional probabilities (see Figure 4.20) produced by either Multivariate Gaussian or the SVC. The model was built from every tenth frame of each video. The state transition matrix (Equation 4.42) is trained during the training phase.

From the transition matrix, we can see that the capsule stays in the small intestine much longer than in the stomach and even longer than in the "entrance" \(a_{12} \gg a_{23} \gg a_{34}\). Since this is a left-to-right model without possibility of skipping states, all the transition matrix elements apart from the main and the first upper diagonal are
Having trained the state transition matrix, the standard Viterbi algorithm (Viterbi, 1967) is used to estimate the most probable sequence of states. Those states are time-aligned with the feature sequence that generated them. Once the transitions between states have been found, the video can be segmented into respective parts corresponding to Entrance, Stomach, Intestine and Colon.

4.7 Single image classification experiments

We were provided with 76 annotated capsule endoscopy videos from the Norfolk & Norwich University Hospital. The videos were annotated by an experienced clinician and segmented into meaningful parts: Entrance, Stomach, Intestine and Colon.

In our first experiment, we tested three single image kNN, SVC and MG classifiers,
namely: Entrance/Stomach, Stomach/Intestine and Intestine/Colon.

From each video we randomly selected 1000 image frames from each region. Next, we trained each classifier using 10000 (2*5000) frames randomly selected from the above image set. For the SVC classifier, we tested two kernels (radial and linear). The former was chosen as it had been reported to be the most successful in many applications, and the latter kernel was tested mostly for comparison purposes with the former. The optimal values of parameters $C$ (both kernels) and $\gamma$ (radial basis kernel only) of the classifier (see Section 4.5.3) were found using a grid search involving ten-fold cross-validation as illustrated in Figure 4.21. The experiments involving SVC were performed using an open-source toolbox (Chang and Lin, 2001).

The $HS$ histograms were quantised into $32 \times 32 = 1024$ bins. LBP histograms were built using 8 sampling points to provide 7 unique patterns, $21(3 \times 7)$ bins for the independent 1D histogram and $343(7^3)$ for the joint 3D histogram. Motion features were extracted as described in Section 4.4.3 using ARPS block matching algorithm with the allowable displacement parameter set to eleven pixels. Features $S_1 - S_6$ were calculated from a window of 41 frames.

The experiment consisted of two stages. In the first stage we tested only feature vectors extracted from entire image frames, in the second we narrowed our experiment to only those image frames where at least one tissue sub-image region (SubIR) could be created (see Section 4.4.4). Consequently, in the second stage, features were extracted from the entire frames (as in the first stage) as well as from SubIR. The aim of this experiment was to find the set of features which yield the best classification results, hence we tested a number of feature vector combinations that are listed below. In all cases features were compressed using PCA (LBP, DFT) and the hybrid transform (HS). The list of feature vectors tested in the first and the second stage of the experiment can be seen in Tables 4.3 and 4.4.
### Table 4.3: List of entire image feature vectors tested in the first stage

<table>
<thead>
<tr>
<th>Feature vector</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. HS</td>
<td>HS histogram</td>
</tr>
<tr>
<td>2. LBP_3D</td>
<td>3D LBP histogram</td>
</tr>
<tr>
<td>3. LBP_1D</td>
<td>1D LBP histogram</td>
</tr>
<tr>
<td>4. DFT</td>
<td>DFT features</td>
</tr>
<tr>
<td>5. LBP_3D_HS</td>
<td>3D LBP and HS histograms</td>
</tr>
<tr>
<td>6. LBP_3D_HS_DFT</td>
<td>3D LBP, HS histograms and DFT features</td>
</tr>
</tbody>
</table>

### Table 4.4: List of entire image and SubIR feature vectors tested in the second stage

<table>
<thead>
<tr>
<th>Feature vector</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. HS_ent</td>
<td>HS histogram formed from the entire image</td>
</tr>
<tr>
<td>2. HS_reg</td>
<td>HS histogram formed from the SubIR</td>
</tr>
<tr>
<td>3. LBP_ent</td>
<td>3D LBP histogram formed from the entire image</td>
</tr>
<tr>
<td>4. LBP_reg</td>
<td>3D LBP histogram formed from the SubIR</td>
</tr>
<tr>
<td>5. HS_ent_reg</td>
<td>HS histograms formed from 1) the entire image and 2) SubIR</td>
</tr>
<tr>
<td>6. LBP_ent_reg</td>
<td>3D LBP histograms formed from 1) the entire image and 2) SubIR</td>
</tr>
<tr>
<td>7. LBP_reg_HS_ent_reg</td>
<td>3D LBP and HS histograms formed from the SubIR and HS histogram formed from the entire image</td>
</tr>
<tr>
<td>8. LBP_ent_reg_HS_reg</td>
<td>3D LBP and HS histograms formed from the SubIR and 3D LBP histogram formed from the entire image</td>
</tr>
<tr>
<td>9. LBP_ent_reg_HS_ent</td>
<td>3D LBP and HS histograms formed from the entire image and 3D LBP histogram formed from the SubIR</td>
</tr>
<tr>
<td>10. LBP_ent_HS_ent_reg</td>
<td>3D LBP and HS histograms formed from the entire image and HS histogram formed from the SubIR</td>
</tr>
<tr>
<td>11. LBP_ent_reg_HS_ent_reg</td>
<td>3D LBP and HS histograms formed from 1) the entire image and 2) SubIR</td>
</tr>
<tr>
<td>12. LBP_ent_reg_HS_ent_reg_DFT</td>
<td>3D LBP, HS histograms and DFT features formed from 1) the entire image and 2) SubIR</td>
</tr>
</tbody>
</table>

#### 4.7.1 Results

The results of the first and second stage of the classification experiment can be seen in Tables 4.5 and 4.6 respectively. The results clearly confirm our prediction that the Entrance/Stomach classifier performs better than Stomach/Intestine classifier and much better than Intestine/Colon. The better performance of the Entrance/Stomach classifier can be explained by the clear visual difference between these regions. The Entrance consists of images taken in the outside world before the capsule was swallowed, mouth
and oesophagus. Only frames in the oesophagus are unlikely to be misclassified. The worst, but still satisfactory performance was for the Intestine/Colon classifier. This was due to the significant similarity between the tissues around the IV, and the occasional presence of faecal contamination which obstructs the camera view, and varies considerably between videos.

Table 4.5: Percentage of correct classifications - First Stage

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Entrance/Stomach</th>
<th>Stomach/Intestine</th>
<th>Intestine/Colon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>kNN rad.</td>
<td>SVC rad.</td>
<td>MG lin.</td>
</tr>
<tr>
<td>HS</td>
<td>99.1</td>
<td>99.6</td>
<td>94.5</td>
</tr>
<tr>
<td>LBP_3D</td>
<td>99.3</td>
<td>99.7</td>
<td>97.4</td>
</tr>
<tr>
<td>LBP_1D</td>
<td>99.2</td>
<td>99.1</td>
<td>93.5</td>
</tr>
<tr>
<td>DFT</td>
<td>99.2</td>
<td>99.5</td>
<td>98.5</td>
</tr>
<tr>
<td>LBP_3D_HS</td>
<td>99.5</td>
<td>99.8</td>
<td>98.3</td>
</tr>
<tr>
<td>LBP_3D_HS_DFT</td>
<td>99.6</td>
<td>99.9</td>
<td>99.5</td>
</tr>
</tbody>
</table>
Table 4.6: Percentage of correct classifications - Second Stage

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Entrance/Stomach</th>
<th>Stomach/Intestine</th>
<th>Intestine/Colon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>kNN</td>
<td>SVC</td>
<td>MG</td>
</tr>
<tr>
<td>1 HS_ent</td>
<td>99.6</td>
<td>99.8</td>
<td>89.6</td>
</tr>
<tr>
<td>2 HS_reg</td>
<td>98.7</td>
<td>99.7</td>
<td>87.5</td>
</tr>
<tr>
<td>3 LBP_ent</td>
<td>99.6</td>
<td>99.9</td>
<td>97.3</td>
</tr>
<tr>
<td>4 LBP_reg</td>
<td>99.2</td>
<td>99.9</td>
<td>94.4</td>
</tr>
<tr>
<td>5 HS_ent_reg</td>
<td>99.6</td>
<td>99.9</td>
<td>94.5</td>
</tr>
<tr>
<td>6 LBP_ent_reg</td>
<td>99.4</td>
<td>99.9</td>
<td>97.8</td>
</tr>
<tr>
<td>7 LBP_reg_HS_ent_reg</td>
<td>99.6</td>
<td>99.9</td>
<td>97.4</td>
</tr>
<tr>
<td>8 LBP_ent_reg_HS_reg</td>
<td>99.6</td>
<td>99.9</td>
<td>99.0</td>
</tr>
<tr>
<td>9 LBP_ent_reg_HS_ent</td>
<td>99.7</td>
<td>99.9</td>
<td>98.8</td>
</tr>
<tr>
<td>10 LBP_ent_HS_ent_reg</td>
<td>99.6</td>
<td>99.9</td>
<td>99.0</td>
</tr>
<tr>
<td>11 LBP_ent_reg_HS_ent_reg</td>
<td>99.6</td>
<td>99.9</td>
<td>99.4</td>
</tr>
<tr>
<td>12 LBP_ent_reg_HS_ent_reg_DFT</td>
<td>99.7</td>
<td>99.9</td>
<td>99.7</td>
</tr>
</tbody>
</table>
In almost all cases SVC with the radial basis kernel was the most accurate classifier. It was then followed by kNN, which in the vast majority of cases outperformed SVC with the linear kernel. On the other hand, we can see that Gaussian distributions do not provide a good model of our features (DFT in particular) as the MG classifier performed significantly worse than other three classifiers. What is also interesting is the fact that SVC results remained very consistent as each new feature was added. We could clearly see a gradual improvement as we incorporated extra features. In the case of MG, an improvement can still be seen, but the MG failed on some features, in particular on feature vectors that included motion information (DFT).

With regard to the feature selection, the joint LBP histogram performed notably better than the independent LBP for all classifiers. The HS histogram alone was only slightly worse than the joint LBP, apart from the Intestine/Colon classifier where it had an advantage over LBP. We suggest that the reason for this is the presence of food debris, affecting the Intestine/Colon classification, whose texture is irregular. Thus it can be described more accurately by the colour information alone, as contained in the HS histogram. The best results were obtained when the joint LBP histogram was combined with the HS histogram. This shows us that although LBP carries colour and texture information together, adding the HS histogram boosts the results. This can be easily explained by the nature of the LBP histogram - it carries colour information about edges, but not about the actual distribution of colours.

DFT features alone provide worse classification results than any other single feature classifier. However, when combined with other colour and texture features, they provide the most discriminative feature set (see the last rows of Tables 4.5 and 4.6). These results are not a surprise since we expected motion features to carry less useful information than colour and texture. However, since motion information is independent of colour and texture, combining it with the latter slightly improves the classification results.
The feature vectors incorporating the information taken just from the sub-image regions do not deliver the same accuracy, but when combined with the feature vector containing the knowledge of the entire image, the classification accuracy increases. The most accurate results were achieved when the features of the entire image were combined with those of the SubIR as the results in bold in Table 4.6 show.

### 4.8 Video segmentation experiments

In the second experiment, we tested the performance of different video segmentation methods. We divided this experiment into two stages.

In the first stage, we used only the feature vectors extracted from the entire images (see Table 4.7) and tested these on all types of discriminators (naïve, sliding window and HMM). We used the classifier with the best performance on single images - SVC with radial basis kernel, as a tool providing the input for the discriminator. Each classifier was trained using frames from all videos (except the one to be tested on, which provided 76-fold validation). Next, we carried out the video segmentation as described in the appropriate subsections of Section 4.6.

In case of HMM, we classified each tenth image frame of the video using these classifiers, calculated the conditional probabilities and computed the best possible sequence of states using Viterbi algorithm.

As to the sliding window segmentation method, we tested two different sets of window sizes to illustrate the importance of their choice. The size of the sliding window for the entrance/stomach classifier in both sets was the same - 50 frames. Regarding the remaining two classifiers, the window sizes in the first set were 200 and 500 frames and in the second set 1,200 and 3,000 frames, for stomach/intestine and intestine/colon classifiers respectively. The increasing sizes of sliding windows for subsequent classifiers are motivated by our expectation of the increasing error in the discrimination of
subsequent regions (Tables 4.5 and 4.6 show that, the accuracy of the entrance/stomach classifier is higher than the accuracy of the stomach/intestine and much higher than that of the intestine/colon classifier). To make the results consistent with those obtained for HMM, we also classified only every tenth frame from each WCE video.

With regard to the naïve segmentation method, we ran the algorithm 25 times on each video and compared the result from a single pass of the algorithm, with the median result of 25 passes. In each of the passes, the starting position of region discrimination was chosen randomly, in the vicinity behind the transition point of the regions found by the previous discriminator. For example, for the entrance/stomach discriminator, we randomly select a frame from the first 100 frames of the video, run the search algorithm which returns frame $x$ as the position of the pylorus, then we randomly select a frame from 1000 frames immediately behind the frame $x$, run the stomach/intestine discriminator and repeat the same procedure for the intestine/colon discriminator. The initial steps of the discriminators were chosen to be 32, 1024 and 2048 frames, for entrance/stomach, stomach/intestine and intestine/colon discriminators respectively.

Table 4.7: List of discriminators tested in the first stage of the video segmentation experiment

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HS histogram formed from the entire image (HS_ent)</td>
</tr>
<tr>
<td>2</td>
<td>3D LBP histogram formed from the entire image (LBP_ent)</td>
</tr>
<tr>
<td>3</td>
<td>DFT features (DFT)</td>
</tr>
</tbody>
</table>

In the second stage, we tested only the most advanced of the classifiers that utilise HMM. Apart from the feature vectors tested in the first stage, we also experimented with features calculated from sub-image regions as well as with different combinations of features. As in the first stage, we used the SVC with the radial basis kernel, but in addition we also tested the Multivariate Gaussian, since this classifier provides a classification output in terms of probability, which is later used by the Viterbi algorithm. We trained the set of SVC and MG classifiers using the same procedure as in the first stage.
With regard to discriminators using features extracted from both entire images and SubIRs, the classification sequences consist of images of which only some contain SubIR. Indeed, within each video, the sequence of images lacking SubIR can be notably long (see Table 4.1). This means that not all video frames can be classified with the classifier taking into account SubIR features. However, such frames can still be classified using the classifier based on the entire image features. Therefore, each of the discriminators, which takes into account SubIR features, is built from two classifiers, one classifies only those image frames containing regions of clearly visible tissue and the other the remainder of the image frames as shown in Figure 4.22. As can be seen in Table 4.8, discriminators 4, 5, 7 and 9 take into account SubIR features and hence they are built from two classifiers, whereas discriminators 6 and 8 as well as 1, 2 and 3 from Table 4.7 are built from one classifier which classifies each image frame from the classification sequence based on entire image features only.

**Figure 4.22:** A flowchart showing how discriminators 2, 4 and 6 choose the appropriate classifier for each frame from the classification sequence.
CHAPTER 4. TOPOGRAPHIC VIDEO SEGMENTATION

Table 4.8: Additional feature vectors tested in the second stage of the video segmentation experiment

<table>
<thead>
<tr>
<th>Feature Vector</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>HS histograms formed from 1) the entire image and 2) SubIR (classifies SubIR images) + HS histogram formed from the entire image (classifies the remaining images) (HS_ent_reg__HS_ent)</td>
</tr>
<tr>
<td>5</td>
<td>3D LBP histograms formed from 1) the entire image and 2) SubIR + 3D LBP histogram formed from the entire image (LBP_ent_reg__LBP_ent)</td>
</tr>
<tr>
<td>6</td>
<td>3D LBP and HS histograms formed from the entire image (LBP_ent_HS_ent)</td>
</tr>
<tr>
<td>7</td>
<td>3D LBP and HS histograms formed from 1) the entire image and 2) SubIR + 3D LBP and HS histograms formed from the entire image (LBP_ent_reg_HS_ent_reg__LBP_ent_HS_ent)</td>
</tr>
<tr>
<td>8</td>
<td>3D LBP and HS histograms and DFT features formed from the entire image (LBP_ent_HS_ent_DFT)</td>
</tr>
<tr>
<td>9</td>
<td>3D LBP and HS histograms and DFT features formed from 1) the entire image and 2) SubIR (not DFT) + 3D LBP and HS histograms and DFT features formed from the entire image (LBP_ent_reg_HS_ent_reg_DFT__LBP_ent_HS_ent_DFT)</td>
</tr>
</tbody>
</table>

4.8.1 Results

The accuracy of the video segmentation algorithms was assessed as the frame difference (error) between the point in the video where the boundary has been manually annotated (by a clinician) and the point selected by our algorithm. Tables 4.9 and 4.10 show the mean and the median error of the discriminators tested in the first and the second stage respectively. We consider median to be a better measure of the algorithm performance as the error distributions are highly skewed. When the algorithm fails, sometimes it fails completely and the transition between two regions is placed thousands of frames from the actual location. Such outliers influence the mean of the error and therefore when comparing different methods, we will concentrate on the median.

We have also performed the number of Wilcoxon signed rank tests (WSRT) in order to establish the statistical significance of differences in the performance of different methods. Due to the volume of statistics involved, they can be found separately in Appendix A. In that appendix, we also describe in detail the procedure of the Wilcoxon
Table 4.9: The median and mean errors (in frames) of HMM discriminators tested in the first stage. Note: the entire video length is around 50,000 frames

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>Esogastric Junction</th>
<th>Pylorus</th>
<th>Ileoceleal Valve</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>naive - 1 pass</td>
<td>naive - 25 passes</td>
<td>SW:50, 200, 500</td>
</tr>
<tr>
<td></td>
<td>median   mean</td>
<td>median   mean</td>
<td>median   mean</td>
</tr>
<tr>
<td>HS_ent</td>
<td>36       70</td>
<td>35       70</td>
<td>27       59</td>
</tr>
<tr>
<td>LBP_ent</td>
<td>4        46</td>
<td>3        41</td>
<td>9        46</td>
</tr>
<tr>
<td>DFT</td>
<td>40       75</td>
<td>40       70</td>
<td>14       48</td>
</tr>
<tr>
<td></td>
<td>197      1,131</td>
<td>119      979</td>
<td>134      739</td>
</tr>
<tr>
<td></td>
<td>206      1,126</td>
<td>150      786</td>
<td>135      646</td>
</tr>
<tr>
<td></td>
<td>1,352    3,052</td>
<td>751      2,065</td>
<td>560      1,993</td>
</tr>
<tr>
<td></td>
<td>6,838    7,289</td>
<td>2,489    3,813</td>
<td>965      4,323</td>
</tr>
<tr>
<td></td>
<td>5,908    6,776</td>
<td>2,701    4,356</td>
<td>738      4,110</td>
</tr>
<tr>
<td></td>
<td>4,683    8,620</td>
<td>2,127    4,088</td>
<td>12,013   13,143</td>
</tr>
</tbody>
</table>

As to the first stage of the video segmentation experiment, a few obvious relationships can be spotted. First, as the results of the single image classification experiment suggested, the video segmentation error is smallest for the entrance/stomach discriminator, followed by the stomach/intestine and finally the intestine/colon. Apart from the entrance/stomach discriminator, the simple naïve discriminator with one pass performed notably worse than all the other segmentation methods. The reason for its good performance in the task of Esogastric junction (EJ) location is that for the sliding window and HMM method we used only every tenth frame of the videos, whereas the naïve method terminates when the step is reduced to one frame. We can infer from this that the location of the EJ obtained using the two former methods can be further tuned using the naïve (or in fact sliding window) method in a small window around the initial result.

The results obtained using the multiple passes of the naïve method also compare favourably with the results of the single pass. Generally this improvement can be considered statistically significant which can be seen in Figures A.1, A.2 and A.3. The only discriminators here which did not provide a statistically significant improvement
were the entrance/stomach and stomach/intestine discriminators utilising the HS feature vector (see Figure A.2).

As to the sliding window discriminators, we can see an advantage of using wider windows for stomach/intestine and intestine/colon discriminators. In all six cases from Table 4.9 (two discriminators × three feature vectors), the medians as well as means of the error distributions were lower for these cases where, a wider sliding window was used.

We can also see an improvement of sliding window results over those obtained for the 25 passes of the naïve algorithm. Although this relationship is clear from the data contained in Table 4.9, the Wilcoxon signed rank test generally did not find it statistically significant (at the 0.1 level). In Section A.2.2, four sample statistical tests have been performed. We can see that by increasing the window sizes, the advantage of the sliding window over the naïve method rises (P-values decrease), but is still not statistically significant at the 0.1 level.

As to the HMM method, we can only see an improvement over the sliding window method with the smaller window sizes. There is little improvement when we compare this method with the more accurate sliding window algorithm with wider windows. This is confirmed by a few sample tests, which can be found in Section A.2.3. The lack of significant improvement of HMM over sliding window can be explained by the nature of the state transition matrix (see Equation 4.42). Since all the diagonal elements are close to one and only forward transitions can happen, this does not provide a significant advantage over the sliding window with a large enough window and a predefined order of states (GI regions). The fact that the HMM uses probabilities as the input for the video segmentation stage, as the results suggest, does not yield a large reduction of error compared to the sliding window method, which uses only the sequence of thresholded classification results.

With regard to the performance of different features, we can clearly see that the
DFT features are no match for any of the two colour and texture features. For the best segmentation methods (HMM or sliding window with a larger window), the motion features provide reasonable results for entrance/stomach classification, but for the other two discriminators the median error is around three times higher than for any of the two other colour/texture features. Having analysed the results of the single image classification experiment, this result is not surprising, though.

As to the features based on Hue Saturation and Local Binary Pattern histograms, we can see some advantage of the latter method. For all the video segmentation methods LBP performed better for the entrance/stomach discriminator. It was also superior in most of the segmentation methods for stomach/intestine and intestine/colon discriminator. Figure A.23 illustrates the signed rank test comparing these two feature vectors for the HMM segmentation method. The improvement is statistically significant for both entrance/stomach and stomach/intestine and for the intestine/colon, there is also some improvement, although the P-value = 0.12, which is above the level we consider statistically significant (0.1).

With regard to the second stage of the experiment, whose results we can see in Table 4.10, we can clearly see that our results are consistent with those obtained in the first experiment as in the vast majority of cases, the Support Vector Classifier with the radial basis kernel produced a far smaller error than the Multivariate Gaussian classifier. Usually, this improvement is statistically significant, which can be seen in figures located in Section A.2.4.
Table 4.10: The median and mean errors of HMM discriminators in frames. Note: entire video length is around 50,000 frames

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>Oesogastric Junction</th>
<th>Pylorus</th>
<th>Ileocecal Valve</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>median</td>
<td>mean</td>
<td>median</td>
</tr>
<tr>
<td>HS_ent</td>
<td>SVC</td>
<td>MG</td>
<td>SVC</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>8</td>
<td>56</td>
</tr>
<tr>
<td>HS_ent_reg___HS_ent</td>
<td>24</td>
<td>14</td>
<td>58</td>
</tr>
<tr>
<td>LBP_ent</td>
<td>10</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBP_ent_reg___LBP_ent</td>
<td>9</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBP_ent_HS_ent</td>
<td>9</td>
<td>8</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBP_ent_reg_HS_ent_reg___LBP_ent_HS_ent</td>
<td>9</td>
<td>6</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFT</td>
<td>14</td>
<td>18</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBP_ent_HS_ent_DFT</td>
<td>9</td>
<td>14</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBP_ent_reg_HS_ent_reg_DFT___LBP_ent_HS_ent_DFT</td>
<td>8</td>
<td>6</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Benchmark (see Conclusions) 2 18 279
Regarding discriminators employing features extracted from the entire images as well as SubIRs, we can see a consistent improvement only for stomach/intestine discriminators. This can be seen in all four figures located in Section A.2.5. In two of them the improvement was statistically significant, whereas in other two the P-values of the WSRT were 0.17 and 0.11. Table 4.1 gives us a clue for an explanation of this interesting observation. It tells us that the images with SubIRs are far more numerous in the stomach and the intestine regions. From this we can infer that including additional information from SubIRs will improve most of all the localisation of the Pylorus, as this is the transition point between the two GI regions, which contain most of the SubIRs. In theory, entrance/stomach and intestine/colon discrimination should be improved too, but the results do not reflect this, which can be explained by the smaller number of SubIRs in both the entrance and colon regions.

Combining different types of features usually reduces the median errors, although the improvement is rarely statistically significant (see Figures A.25-A.27).

### 4.9 Prototype

The research described in this chapter resulted in a prototype of the topographic video segmentation software. The software, which can seen in Figure 4.23 was deployed in clinical trials in the Gastroenterology Unit of Norfolk and Norwich University Hospital, Norwich, UK. The software provides a clinician with a graphic user interface allowing him to execute certain capsule video segmentation algorithms, view their results and enter their own annotations. The GUI allows a clinician to choose a naïve or the sliding window method, carry out the video segmentation and compare the result with their own annotation.
4.10 Conclusions

Results of the discriminator experiments are generally consistent with those obtained for the respective classifiers. SVC with the radial basis kernel is the best choice for the classifier. The most successful segmentation methods are HMM and sliding window (given a large enough window size). It can be seen that the best performance is achieved by classifiers which combine features from the whole image and sub-image regions, in particular for the SVC based discriminator. As expected, the localisation of the esogastric junction is the most accurate (the median error of a few frames is negligible since the video was sampled every tenth frame), followed by pylorus. With regard to the IV location, the results deteriorate, but are still satisfactory if we consider
that in many cases it is impossible to know (even for an experienced capsule viewer) exactly when the capsule passes from the intestine to the colon due to severe visual contamination by food debris. It is even more apparent if we compare these results with the results from the bottom row (Benchmark) of Table 4.10, which quotes the figures of an experiment described in (Soares et al., 2006) involving three experienced capsule specialists manually segmenting a set of 47 capsule exams. They show that the manual annotations of these clinicians were quite unstable and represented the same tendency, which can be seen in our results i.e. the localisation of the esogastric junction is the most accurate, followed by pylorus and IV with the biggest error. Even though, our results were not as good as the Benchmark provided by the human experts, they are still sufficiently accurate to enable the clinician to interactively locate a boundary within a few seconds. Indeed, the stomach/intestine discriminator error of 100 frames constitute only four seconds of fast-forward viewing and the error of 500 frames of the intestine/colon discriminator, 20 seconds viewing time. These times are based on the clinician watching the video at 25 frames per second and would vary with viewing speed, which is usually lower. Nevertheless, we have shown that such tools can shorten the time required to annotate GI transition points, and hence significantly reduce the video viewing time.
Chapter 5

Bleeding detection

Another area of Wireless Capsule Endoscopy video analysis that has received considerable interest in the computer vision and image processing community is *Bleeding Detection*. Detection of bleeding can be seen as the next natural step after topographic video segmentation (described in the previous chapter). The reason for this is that we use the video segmentation algorithms as the preprocessing step. Moreover, bleeding events can be considered a subset of the set of all abnormalities in the WCE video. They are also one of the most frequent abnormalities in these videos. Therefore, one can think of the bleeding detection problem as the prelude to the ultimate problem of detecting any abnormality.

In the next section, we consider the motivation behind this work. This is followed by the description of related work (see Section 5.2). In Section 5.3 we give an overview of our system. Next, we propose and describe in detail our novel method of bleeding detection (see Sections 5.4 and 5.5). Then, we describe a software prototype incorporating the algorithms introduced in this chapter (see Section 5.6). The last section (5.7) of this chapter contains the conclusions.
5.1 Motivation

The manufacturers of the Given capsule system provide only one automatic image analysis function with their Rapid Reader software: The Suspected Blood Indicator (SBI), which is designed to report the location of video frames containing active bleeding (see Figure 5.1). However, this tool has been reported to have insufficient specificity and sensitivity (Signorelli et al., 2005). Therefore it does not free the specialist from reviewing the entire footage and should only be used as a fast screening tool.

![Figure 5.1: WCE images containing different manifestations of bleeding.](image)

In (Zanati et al., 2004), the authors report that SBI identified 18 out of 22 (82%) regions distal to the duodenum with visible blood. The SBI was positive on 59 occasions (18 true positives, 41 false positives) in the 42 patients studied, an average of 1 false
positive per patient. The SBI detected 36 of 76 regions with either blood or a red mucosal lesion. It would not detect bleeding lesions in the stomach or altered (mixed, diluted) blood anywhere in the GI tract.

In (D’Halluin et al., 2005), the authors report the sensitivity and specificity of 83% and 66% for the presence of fresh blood and 37% and 59% for the presence of other lesions. The authors conclude that SBI detection of intestinal lesions is of limited clinical value and does not reduce the time required for interpretation of the capsule endoscopy procedure.

To date, we have not been able to find any publications evaluating the red spot detection tool in the Olympus EndoCapsule.

These studies suggest that there is a need for improvement in terms of detection rates of bleeding areas and other red lesions and this need is the most important motivation for the research described in this chapter.

5.2 Related work

It is not obvious how exactly SBI detects red areas since Given Imaging has not revealed this. Similarly, we do not know how Olympus EndoCapsule red spot detector works. In (Signorelli et al., 2005), the authors state that in their observations the ability of SBI to detect red lesions was related neither to their size nor to the length of the lesion event (number of frames).

In (Hwang et al., 2006; Cox, 2005), the authors propose a new algorithm that as they claim can detect bleeding areas in the capsule videos. The algorithm uses Expectation Maximization (EM) clustering and Bayesian Information Criterion (BIC). The authors manually segmented around 200 images into blood and non-blood regions. Then, they selected 16,000 bleeding and 45,000 non-bleeding pixels and modelled the colour distribution of these regions using Gaussian mixtures in RGB colour space. Bayesian
CHAPTER 5. BLEEDING DETECTION

Information Criterion was used to decide the number of clusters. In the first step of the algorithm, dark pixels are removed. In the second step, the algorithm chooses those pixels \( x \) to be bleeding candidates for which conditional probability \( p(x|\text{bleeding}) \) of a pixel \( x \) given by bleeding pixels is significantly larger than conditional probability \( p(x|\text{non-bleeding}) \) of a pixel \( x \) given by non-bleeding pixels; and also it is larger than a certain predefined threshold. In the final step of the algorithm, the areas of bleeding regions are calculated and all segmented regions containing less than 1,000 pixels are rejected. To test the results of bleeding detection, the authors selected 15,222 capsule images of which 1,731 contained blood from three different videos. On this test set, a reported specificity and sensitivity were 98.10% and 92.55% respectively. Unfortunately, this algorithm was not tested on full-length video sequences, which makes it difficult to state whether it performs better than SBI.

5.3 Method overview

Bleeding detection in capsule endoscopy exams must take into account several issues: 1) time-varying intensity of the illuminating light source; 2) changes in colour of bleeding areas due to alteration by gastrointestinal liquids; 3) detection of non-bleeding lesions which still manifest themselves by more reddish and pinkish appearance than the surrounding healthy tissue; 4) capsule position with respect to gastrointestinal organs (oesophagus, stomach, intestine, colon); detecting when the capsule leaves the stomach is particularly important since the stomach tissue is (as was noted in Chapter 4, see Figure 4.3) more reddish than the intestine, which will affect the detection of any red objects.

Statistical colour models have been widely used in applications involving detection and tracking (Celik et al., 2006). Particularly, many papers have been published in the area of skin colour modelling (Jones and Rehg, 1998; Bosson et al., 2002; Sigal et al.,
since skin segmentation is used as a preprocessing step in applications such as face and gesture recognition, blocking pornography etc. These statistical models are either parametric or histogram based.

Our approach uses adaptive colour histograms to track a moving background and bleeding colour distributions over time. Such an approach addresses the problem of drastic changes in blood colour distribution when it is altered by gastrointestinal fluids and allow detection of other red lesions, which although usually "less red" than fresh active bleeding, can still be detected when the difference between their colour distributions and the background is large enough. Contrary to (Hwang et al., 2006; Cox, 2005), where the authors use parametric bleeding colour distribution modelling, we have chosen a histogram based method. This was motivated by the need of fast model adaptation, which is easier in the non-parametric method.

The pylorus (the valve between the stomach and the intestine) is detected using the algorithm described in Chapter 4. Since we had no video data containing bleeding in the stomach, we concentrated on the analysis of the video frames after the capsule has passed into the intestine. It is worth noting that the SBI function in the Given Pillcam software prompts the gastroenterologist to manually annotate this point in the footage in order to display the bleeding detection results. Incorrect placement of pylorus annotation before its actual position often results in numerous false positives in the stomach, which as was mentioned above is due to the fact that stomach tissue color distribution is more reddish than the intestine and hence more similar to the blood colour distribution.

Figure 5.2 contains a simplified flowchart of our bleeding detection system. First, we apply a pixel classifier to each pixel of the image being classified (for details, see Section 5.4). Then, we perform a region growing operation in which candidate pixels are merged into regions of at least 250 pixels. If a blood region is detected, we extract from it colour and texture features. These features are also extracted from the region
surrounding the suspicious region. Additionally we also search for specular highlights in the image in order to check whether it contains air bubbles, which, as explained in Section 5.5.1, can cause false positives. All these features are used in the last step which involves the classification of all the images, that the pixel classifier has labelled as containing suspicious regions (see Section 5.5). The images are classified using a Support Vector Classifier into three classes: Bleeding, Lesion/Abnormality or Normal.

Figure 5.2: Bleeding detection system

5.4 Adaptive histogram based pixel classifier

In the following sections we describe the details of the pixel classifier being the first part of our method.

5.4.1 Colour space choice, Prior histogram learning

The HSI colour space was chosen to model the bleeding and non-bleeding colour distributions. The range of colours present in WCE images is relatively small, mapping
to a region covering just around 20% of the possible $HS$ colour space, and so we equalise the histograms to this subset (of red to yellowish-green) colours. This approach has been successfully used in our capsule video segmentation algorithm, as described in Chapter 4. The $HSI$ histograms were quantised to $32 \times 32 \times 24$. The bleeding histogram was built from $\sim 100,000$ pixels segmented manually from 40 WCE images. The non-bleeding histogram was built from a significantly larger set of 8,400 images selected from the set of 84 full length videos available to us from the Norfolk and Norwich University Hospital. Figures 5.3 and 5.4 show bleeding and non-bleeding colour distributions respectively. These figures contain only 15 intensity slices (4-18), which contain most of the pixel count.

**Figure 5.3:** Bleeding 3-D HSI histogram sliced along different intensity levels

### 5.4.2 Histogram Adaptation

Modelling "background" in capsule examinations is somewhat different to other applications such as skin detection. On the one hand we want to adapt the background colour distribution to the current place in the footage, but on the other hand a capsule
endoscope moves in a very constrained environment. This is the reason why we want to retain the memory of the background, as learned from the training set. The flowchart describing the dynamic updates of histograms can be seen in Figure 5.5.

The formula for background histogram updates is given by Equation 5.1. The first term of this equation (controlled by the parameter $\gamma$) allows us to hold to certain extent to the original training set (denoted by $H_{NB}^{org}$) and the second term (controlled by the parameter $\alpha$) is responsible for the speed of histogram adaptation. The histograms are normalised before each update.

$$H_{NB}(t) = (1 - \gamma) \times H_{NB}^{org} + \gamma \times [(1 - \alpha) \times H_{NB}(t - 1) + \alpha \times H_{NB}(t)] \quad (5.1)$$

The adaptation of the bleeding colour distribution is less important than the background adaptation since the colour of blood varies far less than the background. Moreover, we are particularly interested in detecting blood events and pointing them to the

---

Figure 5.4: Non-bleeding 3-D HSI histogram sliced along different intensity levels
Figure 5.5: The flowchart illustrating how the dynamic updates of the histograms are performed.
clinician since they can later view the adjacent frames manually. However, we are still using the same adaptation method for the foreground (bleeding) since this should allow us to track each bleeding event more precisely (see Equation 5.2). In this equation, we denote the adaptation parameters $\beta$ and $\delta$; analogically $H^\text{org}_B$ is the bleeding colour distribution learned as described in the previous section.

$$H_B(t) = (1 - \delta) \ast H^\text{org}_B + \delta \ast [(1 - \beta) \ast H_B(t - 1) + \beta \ast H_B(t)] \quad (5.2)$$

5.4.3 Pixel classification

A pixel is considered to be a candidate bleeding pixel if the two following conditions are met: first, the conditional probability of a pixel $x$ given by bleeding pixels divided by the conditional probability of a pixel $x$ given by non-bleeding pixels is larger than the certain threshold $T_1$: $\ln \frac{p(x|\text{bleeding})}{p(x|\text{non-bleeding})} > T_1$; and second the conditional probability of a pixel $x$ given by bleeding pixels is larger than the threshold $T_2$: $\ln p(x|\text{bleeding}) > T_2$. These thresholds as well as adaptation parameters are learned in the first experiment described in Section 5.4.4.

5.4.4 Finding optimal adaptation coefficients

In the first experiment, the adaptive pixel classifier was tested on 23 video sequences, each containing 100 frames extracted from different capsule exams.

Receiver Operator Curves

To present the results of the first experiment, Receiver Operator Curves (ROC) have been used (Webb, 2002; Alexander, 1997). They are commonly used in signal detection, which is a binary decision making problem. There are two inputs: positive and negative; and the system can classify them correctly or incorrectly, giving rise to four
possible decision outcomes: true positive, if a signal is correctly classified as positive; false positive if it is wrongly classified as positive. A case that is correctly classified as negative is called true negative; whereas cases that are incorrectly classified as negatives are called false negatives. Our assessment uses true positive and true negative ratios, which can be computed by dividing the number of true positives by the actual number of positives (\(TPr = \frac{TP}{TP+FN}\)) and the number of true negatives by the actual number of negatives respectively \(TNr = \frac{TN}{TN+FP}\). The former ratio is also often referred to as sensitivity or recall and the latter as specificity. The ROC curve is obtained by plotting those ratios against each other, with a varying parameter (in our case different adaptation parameters). The ideal system operating point lies in the upper-left corner \((TPr = 1, 1 - TNr = 0)\) of the plot.

F measure

Another way of presenting the results involves the use of F measure (van Rijsbergen, 1979), which is an explicit or composite measure of effectiveness i.e. it takes into account two basic measures of classification effectiveness: recall \(R\) and precision \(P\). The former is equal to true positive ratio \(TPr\), which was introduced earlier, whereas the latter equals the ratio of relevant samples retrieved to the number of all retrieved samples \(P = \frac{TP}{TP+FP}\).

\[
F = \frac{(\beta_F^2 + 1) \times P \times R}{\beta_F^2 \times P + R}
\]  

(5.3)

where \(\beta_F\) denotes the relative importance of precision and recall i.e. if we attach no importance to the precision then \(\beta_F \rightarrow \infty\) and in the opposite situation when we attach no importance to the recall then \(\beta_F \rightarrow 0\). In our case, we used a balanced value \(\beta_F = 1\) which weights precision and recall equally. Such a composite measure has an advantage of describing performance by only one factor and hence it eases comparisons of different methods or parameters.
CHAPTER 5. BLEEDING DETECTION

Experiment

In order to evaluate the results all the frames were annotated according to the bleeding contents they carry. We had to decide with what labels the frames should be annotated. From our first observations it was apparent that there are different degrees of "bloodiness" in the capsule exams, which influence the effectiveness of any bleeding detection system. After some consideration, it was decided to annotate bleeding frames with two labels: 1) clear fresh bleeding, where the clinician only needed to view a single frame from the event, in order to state categorically that this was actual bleeding; 2) altered blood, blood often mixed with intestinal fluids, where viewing only a single frame was insufficient for the clinician to state categorically that this was a frame containing a form of bleeding. Moreover, a separate label was given to all the other lesions, which were not related to bleeding, but manifested themselves by different degrees of redness.

Next, we ran our classification algorithm on these 23 videos for varying parameters of $T_1$, $T_2$, $\alpha$, $\beta$, $\gamma$ and $\delta$. We used only the images with the first label (clear fresh bleeding) for calculation of $TPr$, $TNr$ and Precision i.e detection of blood in frames labelled as altered blood was considered as true positive, but not detecting blood in such frames was not considered false negative, but true negative. This was motivated by the fact that after initial observations, we realised that treating altered blood, which is very similar to the normal tissue, the same way as clear fresh blood would produce too many false positives. The ROC and $F$ measure curves were used to find the optimum values of adaptation parameters, which was found to be $T_1 = 3$, $T_2 = -10$, $\alpha = 0.1$, $\beta = 0.2$, $\gamma = 0.95$ and $\delta = 0.2$.

Figures 5.6 and 5.7 contain ROC curves for varying adaptation parameters of Equations 5.1 and 5.2 respectively. Figures 5.8 and 5.9 contain the same data which were used to calculate the composite $F$ measure and hence the different curves are easier to compare. From these figures, we can see that adaptation coefficients of background ($\alpha$, $\gamma$) have much higher impact on classification results than adaptation coefficients of
foreground (blood) \((\beta, \delta)\). Moreover, we can see that parameter \(\gamma\) should be close to 1, but if it reaches 1 the performance suddenly drops. This clearly shows us that it is important to retain some information of the original background learned from the training set throughout the entire video scan process.

![ROC curves](image)

**Figure 5.6:** ROC curves for \(T_1 = 3, T_2 = -10\) and \(\beta, \delta = 0\). Different curves represent different \(\gamma\), along each curve increasing \(\alpha\).

### 5.5 Classification of detected regions

As stated earlier in this chapter, close observation of bleeding regions suggests that the appearance of blood in the WCE images can be altered by many factors e.g. intestinal fluids. Moreover many lesions appear redder than the normal tissue and thus, to some extent, they can be detected by the blood detection algorithms, which are sensitive to image redness. The clinician focuses on image features, such as the colour of the observed region, but what attracts his attention the most is the rapid change in the video in both temporal and spatial domain.

The adaptive algorithm presented in the previous sections attempts to detect rapid
Figure 5.7: ROC curves for $T_1 = 3$, $T_2 = -10$ and $\gamma = 0.98$, $\alpha = 0.1$. Different curves represent different $\delta$, along each curve increasing $\beta$.

Figure 5.8: $F$ measure curves for $T_1 = 3$, $T_2 = -10$ and $\beta, \delta = 0$. Different curves represent different $\gamma$. 

changes in the video colour distribution, but it lacks the detailed analysis of the detected group of pixels and its neighbourhood. Therefore, the final step of the bleeding detection must be the classification of the detected regions. This could be carried out in a number of ways e.g. by using simple thresholding of region area (already implemented into the region growing process) or more complicated thresholding of region area with relation to the likelihood of region being blood (thresholds $T_1$ and $T_2$). It is also possible to extract more sophisticated region features and then classify them using one of the classifiers described in Section 4.5. The latter approach uses more information of the image region, thus should provide better classification results and therefore it was chosen to be used in our application. In the following subsections, we describe the region feature extraction process, the classification experiments and finally give the results of our bleeding detection algorithm.

Figure 5.9: $F$ measure curves for $T_1 = 3$, $T_2 = -10$ and $\gamma = 0.98$, $\alpha = 0.1$. Different curves represent different $\delta$. 
CHAPTER 5. BLEEDING DETECTION

5.5.1 Feature extraction

Active bleeding usually has clear edges that separate its region from the surrounding tissue. Moreover certain "reddish" lesions or sometimes even bleeding are in a very close proximity to the yellow-white ulceration. In presence of yellow-green colours, the blood colour distribution tends to be altered towards colours, which are often present in normal WCE frames in shadows created by the intestinal folds or air-bubbles. Therefore, combining texture and colour features of the region and its neighbourhood could give enough information allowing the correct discrimination between the false positive and the actual abnormal region.

Colour

As in Chapter 4, page 42, we use $HSI$ colour space and equalise the histograms to the subset of red to yellowish-green colours, which are the only hues present in normal conditions in capsule videos. Unlike the $HS$ histograms used in Chapter 4, here we form 3-D $HSI$ histograms. Including intensity information in our colour descriptors was motivated by the fact that blood usually appears darker than the surrounding tissue (see Figures 5.3 and 5.1). We used $32 \times 32 \times 24$ sampling, which produces a large 3-D structure requiring further compression. This is achieved by the means of 3-D DCT transform (see Section 4.4.5), which produces a 286 element long vector.

According to this recipe we build two histograms from two different regions (see Figure 5.2). The first region denoted as $BR$ (blood region) is the region returned by the pixel classifier. The second region denoted as $NR$ (neighbourhood region) is created using the morphological dilation operation of the $BR$ with the structuring element being a disk of ten pixel in diameter. The colour distribution extracted from $NR$ carries an important information on the immediate neighbourhood of the suspicious region ($BR$) e.g. whether the $BR$ is surrounded by the intestinal fluids which may have altered the blood colour distribution or whether it contains the colour cues suggesting an ulceration.
Similarly to the $BR$, this histogram is also compressed with the 3-D DCT transform producing the same number of feature vector elements. Hence, the combined colour feature vector has 572 elements.

**Texture**

Region texture should also aid in a process of blood non-blood classification, as the blood region boundary is an important visual feature. The texture feature we use for this task is the 3-D Local Binary Pattern ($LBP$) histogram, which has already been described in detail in Section 4.4.2 and performed well in the topographic video segmentation task described in Chapter 4. As with the $HSI$ histograms, we build the $LBP$ histograms from two regions: $BR$ and $NR$ as defined in the previous sub-section. As in Chapter 4, we chose the number of sampling points to be eight, which produces $343(7^3)$ histogram bins. The points were sampled in a radius of two pixels. Hence, for two regions we have 686 feature vector elements.

The combined $HSI$ and $LBP$ feature vector contains $686 + 572 = 1258$ elements and requires to be compressed for further processing. As in Chapter 4, we perform this task using Principal Component Analysis (PCA, see Section 4.4.5).

**Air bubble detection**

It was found during the experiments that a large number of false positives were due to presence of air bubbles in the GI track. The healthy tissue colour distribution seen through the air bubble is similar to the blood colour distribution. The air bubbles tend to appear suddenly, which is also a factor that can trigger false positives in an adaptive algorithm, such as this one.

In order to circumvent this problem, it is necessary to know whether the candidate blood region is within the air bubble. From the images in Figure 5.10, we can clearly see that the air bubbles usually contain specular highlights on their surface. Detecting
them would signify the high probability of the air bubble around them and would allow us to reject any candidate blood region in their neighbourhood.

Figure 5.10: Two images containing air bubbles with specular highlights.

The method, we are using for detection of specular highlights was introduced by Ortiz and Torres (Ortiz and Torres, 2006) and involves the analysis of the $MS$ Intensity-Saturation histogram of the image ($M$ denotes Intensity as it is the mean of $R$, $G$ and $B$ channels). The value of saturation $s$ is calculated from the equation below:

$$s = \begin{cases} \frac{1}{2}(2r - g - b) = \frac{3}{2}(r - m) & \text{if } (b + r) \geq 2g \\ \frac{1}{2}(r + g - 2b) = \frac{3}{2}(m - b) & \text{if } (b + r) < 2g \end{cases}$$  \hspace{1cm} (5.4)

where $r = \max(R, G, B)$, $g = \text{mid}(R, G, B)$ and $b = \min(R, G, B)$

As in (Ortiz and Torres, 2006), we construct the MS histogram (see Figure 5.11), which is the positive projection of the corners of the $RGB$ cube in a normalization of the achromatic line to the $m$ signal. The MS histogram is a grey-scale image in which each coordinate $(m, s), m \in [0, 255], s \in [0, 255]$ represents the pixel count of values $(m, s)$ in the original colour image. The third column of Figure 5.12 shows MS histograms. It can be seen that the values of $m$ and $s$ in both MS histograms are limited to the silhouette defined in Figure 5.11).

The specularities in the chromatic image have low saturation values $s \in (0, s_{max})$
and high intensity values $m \in (m_{\text{min}}, 255)$. When the intensity decreases, the specular reflection becomes diffuse and its colour becomes similar to the surface colour approaching $c_3$ and $c_4$ lines:

\[
\begin{align*}
    c_3 &= -\frac{3}{2}(m - 255) \quad \forall m \in [m_3, m_{\text{max}}] \\
    c_4 &= -3(m - 255) \quad \forall m \in [m_4, m_{\text{max}}]
\end{align*}
\]

where $m_3 = \frac{2}{3}m_{\text{max}}$ and $m_4 = \frac{5}{6}m_{\text{max}}$.

![Figure 5.11: The shape of the MS diagram](image)

Different images may have different dynamic ranges and therefore, the $m$ and $s$ values may not correspond to the positions previously described in the MS diagram. This problem could be circumvented by histogram equalisation of the $m$ signal, but this can often cause an excessive increase in saturation and intensity, resulting in false detections of objects, which are not specular highlights. Instead this, Ortiz and Torres suggest performing a local contrast enhancement using the morphological top-hat contrast operator.
(Soille, 1997) defined as:

\[ m' = m + WTH(m) - BTH(m) \]  \hspace{1cm} (5.6)

where \( WTH(m) = m - \text{opening}(m) \) and \( BTH(m) = m - \text{closing}(m) \).

Performing this operation results in positioning the specular pixels on \( c_3 \) and \( c_4 \) lines in the MS histogram (see Figure 5.12f). We determined the optimum height of these lines experimentally \( - m_{\text{min}} = 210 \). Any pixels with higher intensity than \( m_{\text{min}} \) that lie on the lines \( c_3 \) or \( c_4 \) are thus, classified as specular.

**Figure 5.12:** Contrast enhancement and specularity detection: a,d) original and contrast enhanced images; c,f) their MS diagrams; and b,e) detected specular pixels
5.5.2 Region classification experiment

Having obtained the adaptation coefficients and thresholds from the first experiment, we can run the algorithm on 84 full length video sequences. Since the data set of our 84 videos did not contain many bleeding events, we ran our algorithm on a further twenty 100 frame videos available from the Given Imaging website. This allowed us to extract more bleeding frames. The suspicious frames detected in this experiment were then further classified using Support Vector Classifier (see Section 4.5.3). In order to perform this task, firstly we had to annotate the frames detected by the adaptive pixel classifier into the several target classes. By looking at the frames detected by the pixel classifier, we decided to split the detected images into four classes: 1) active bleeding (583 cases), 2) lesion(s)/abnormal redness (386 cases) and 3) false detections (1292 cases). From each of these frames, we extracted the HSI and LBP features as described in the previous section. If the frame contained more than one suspected region, the features were extracted from the largest region. Next, these features were further compressed using PCA.

Moreover, we carried out the process of specular highlight detection. To each suspected bleeding region, we assigned a specularity feature $F_s$, such that:

\[
F_s = \begin{cases} 
1 & \text{if } \exists \ p \in S \in BR \\
0.5 & \text{if } \exists \ p \in I \wedge p \notin BR \\
0 & \text{otherwise}
\end{cases}
\]  

(5.7)

where $S$ denotes a set of specular pixels, $BR$ a set of pixels contained in the suspected region and $I$ the set of all pixels in the image being analysed.

In addition we use two more features:

\[
F_1 = \sum_{i \in B} Lik_1(i) - T_1 + Lik_2(i) - T_2
\]  

(5.8)
where $\text{Lik}_1(i) = \ln \frac{p(x_i|\text{bleeding})}{p(x_i|\text{non-bleeding})}$ and $\text{Lik}_2(i) = \ln p(x_i|\text{bleeding})$ and $B$ denotes the largest suspected region. The second feature $F_2$ is calculated analogically with an exception that the summation is over all suspected regions (as opposed to the largest bleeding region) in the image. These two features tell us how much above the thresholds $T_1$ and $T_2$, the appropriate likelihood values of the region pixels are. The last feature is the area $F_A$ of the largest blood region.

We tested a range of feature combinations, which are summarised in Table 5.1.

<table>
<thead>
<tr>
<th>feature vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 $\text{HSI}_B\text{R}$</td>
</tr>
<tr>
<td>2 $\text{HSI}_B\text{R}$ and $\text{HSI}_N\text{R}$</td>
</tr>
<tr>
<td>3 $\text{LBP}_B\text{R}$</td>
</tr>
<tr>
<td>4 $\text{LBP}_B\text{R}$ and $\text{LBP}_N\text{R}$</td>
</tr>
<tr>
<td>5 $\text{HSI}_B\text{R}$ and $\text{LBP}_B\text{R}$</td>
</tr>
<tr>
<td>6 $\text{HSI}_B\text{R}$, $\text{HSI}_N\text{R}$, $\text{LBP}_B\text{R}$ and $\text{LBP}_N\text{R}$</td>
</tr>
<tr>
<td>7 $\text{HSI}_B\text{R}$, $\text{HSI}_N\text{R}$, $\text{LBP}_B\text{R}$, $\text{LBP}_N\text{R}$ and $F_S$</td>
</tr>
<tr>
<td>8 $\text{HSI}_B\text{R}$, $\text{HSI}_N\text{R}$, $\text{LBP}_B\text{R}$, $\text{LBP}_N\text{R}$, $F_S$, $F_1$, $F_2$ and $F_A$</td>
</tr>
</tbody>
</table>

For our experiment, we chose a SVC classifier with radial basis kernel that provided the best results in Chapter 4. Since we have more than two target classes we decided to train and test three different types of classifiers:

- **bleeding vs normal** (BvN) - the second target class contained all the frames labelled as false positives.

- **lesion vs normal** (LvN) - the first target class contained all the frames labelled as lesions, whereas the second class contained the false positives.

- **lesion & bleeding vs normal** (LBvN) - the first target class contained all the frames labelled as lesions or bleeding, whereas the second class contained the false positives.
We decided to use these three classifiers as we are not interested in classifying frames precisely into the three original classes. First, we are interested in establishing, which frames are clear bleeding occurrences (classifier BvR) and second we need to check to see if some of the remaining frames are suspicious (classifier LvN). Such a design also allows us to present to the clinician the clear bleeding images first, followed by other images, which we consider to be worth looking at. The final classifier (LBvN) was added in the later stage, and was motivated by the fact that LvN produced too many false positives. Hence, we decided to use both classifiers and consider the frame to contain a lesion/abnormality only if both classifiers (LvN and LBvN) agreed.

We trained each classifier using randomly selected frames from the set mentioned earlier in this section. The optimal values of parameters $C$ and $\gamma$ of the classifier (see Section 4.5.3) were found using the grid search involving ten-fold cross-validation as illustrated in Figure 5.13.

5.5.3 Results

Figures 5.14, 5.15 and 5.16 show the classification results obtained using ten-fold cross validation for BvN, LvN and BLvN classifiers respectively.

From Figures 5.14-5.16 we can see that the highest accuracy (over 97%) was unsurprisingly achieved by BvN classifier. This can be easily explained by the fact that blood regions are visually more homogenous group than lesions; and therefore the former can be more accurately differed from the normal tissue than the latter. The remaining two classifiers (LvN and LBvN) had slightly lower accuracy ($\sim 92\%$), which we still find satisfactory.

The highest accuracy for each of the feature vectors were achieved in most cases for 32 of 64 principal components. This tells us that there is a lot of redundancy in the $HSI$ and $LBP$ histograms, which can be removed using this kind of compression without a negative impact on performance. We can also see that using larger number
of PCs often slightly reduces the performance of the classifier - information carried by
the higher PCs can be regarded as noise.

The most interesting observation from the classification results is the fact that com-
bining the features extracted from the blood region together with those from the region
neighbourhood significantly boosted the accuracy of the classification (usually by 2-
5%).

With regard to the performance of different feature vectors, we can clearly see that
HSI performed significantly better than LBP. This can be explained by the fact that
not all the pixels are included in the LBP histogram (see Equation 4.4). Combining
the blood region with its neighbourhood also gives poorer results for LBP than HSI.
The LBP feature vector built from both blood region and its neighbourhood performed
worse than HSI feature vector built from blood region only. When combining HSI with

Figure 5.13: A grid illustrating the search process of optimal $C$ and $\gamma$ parameters for a
radial basis SVC involving feature vector no. 8 from the table above.
Figure 5.14: Classification results - blood vs normal. (BvR)
the LBP features, in most of the cases we can see some improvement, which is however not high (1%). Extending HSI and LBP by adding $F_s$ (specularity) feature improves the results just slightly and only for BvN and LBvN classifiers. We can explain this by the fact that bleeding regions very rarely contain specular pixels, whereas in some lesions or in their neighbourhoods they are more common. The best accuracy in all cases is provided by the feature vector combining all the colour and texture features as well as additional features $F_s$, $F_1$, $F_2$ and $F_A$. However, for the LvN classifier this result is hardly better than that for the feature vector combining all the colour and texture features. This tells us that not only the specular pixels in the lesion regions, but also features $F_1$ and $F_2$ play a smaller part in the discrimination between the lesion and the normal tissue, as the likelihoods $Lik_1$ and $Lik_2$ depend on region similarity to the original bleeding training set - in other words they depend on region "bloodiness" or "redness". Some lesions may have only small reddish part (low $Lik_1$ and $Lik_2$), that allowed the pixel classifier to detect them, but are then filtered from among the normal cases by the region classifier, using e.g. information contained in the region neighbourhood (e.g. ulceration case). The region size $F_A$ may also have a smaller impact on lesion/abnormality detection as these image features are much more varied in size than bleeding regions.

In this experiment we investigated different feature vectors of three different classifiers. We have also learned the optimal settings in terms of number of principal components as well as parameters $C$ and $\gamma$ of the SVC classifier. In the next subsection, we will use this in an experiment in which we compare the performance of our method with the Given Imaging Suspected Blood Indicator.

5.5.4 Comparison with Given SBI

This experiment involved carrying out the comprehensive comparison of our algorithm with the Given Suspected Blood Indicator. Here, we used only the best performing
Figure 5.15: Classification results - lesion vs normal (LvN)
Figure 5.16: Classification results - blood & lesion vs normal (LBvN)
CHAPTER 5. BLEEDING DETECTION

feature vector combining all the features discussed in the previous section (row eight in the Table 5.1). We trained three previously described classifiers: BvN, LvN and LBvN using the parameters $C$ and $\gamma$ learned in the previous experiment. This time, we performed 84-fold cross validation i.e. for each of 84 videos we trained a different triple of classifiers, which were trained using the training set obtained from all, but one video (the one to be tested on), 83 full-length videos from NNUH plus twenty 100 frame videos from the Given Imaging website).

We used the strategy of combining the detected frames into events so that any two frames separated by no more than ten frames will be stitched into one event. Table 5.2 contains the comparison on per video basis of the performance of SBI and our method.

Calculation of a simple performance measure, such as recall (sensitivity) is impossible here, because there are too many individual frames to annotate (each video consists of $\sim 50,000$ frames). Moreover, sometimes the clear distinction between the abnormal video region and the normal tissue is not clear, which makes annotating every frame in the video even more difficult. Despite this, we can compare the sensitivities of the two methods in relation to one another. In the case of bleeding, SBI has detected total of 18 events comprising 136 frames, whereas our method detected 130 events comprising 814 frames (see last row of Table 5.2). Of even more importance, our method detected bleeding in three videos, which SBI missed completely. From the point of view of the clinical diagnosis, this is far more important than the comparison of the number of frames detected by both methods as the ideal algorithm should be capable of detecting all bleeding events i.e. avoiding any false negatives (Type II error) is more important than avoiding false positives (Type I error).

With regard to lesion/abnormality detection, SBI detected 124 events comprising 681 frames. For our method, these figures are 1,344 and 9,769 respectively. Hence, as with bleeding, we can infer that the sensitivity of our algorithm is more than ten times higher than SBI. As to the number of videos involved, SBI detected lesions/abnormalities
in 14 videos as compared to 37 videos in which our algorithm produced detections.

It is important to establish whether the increased sensitivity of our method was caused by the loss in precision. The number of false positives for SBI is indeed lower than for our method: 24 and 136 as compared to 191 and 345. We calculate two types of precision: 1) for events $P^e$: as the number of relevant events (bleeding and lesion) to the number of all events retrieved (bleeding, lesion and false positives) and 2) for frames $P^f$: as the number of relevant frames to the number of all frames retrieved. For these calculations, we used the values from the last row of Table 5.2. According to this, $P^f_{SBI} = 95.8$, $P^e_{SBI} = 85.5$, $P^f_{our} = 96.6$ and $P^e_{our} = 88.5$. Thus, we can see that the increased sensitivity of our method did not caused the loss in precision. On the contrary, the precision of our new method is slightly higher than that for SBI.

It is also interesting to note that false positives (long sequences of FPs in particular) tend to occur in videos where there are other relevant images detected (see videos no. 5, 24, 26 39, 40, 44, 51, 63, 71, 72, 83 in Table 5.2). Their location in the video tends to be related to the vicinity of the bleeding/lesion events. This suggests that some of those FPs, although visually appearing normal, may not be false positives after all; and may signify some underlying abnormality, which is difficult to spot by the clinician without a wider context of the overall video. The presence of such false positives can be also explained by the way how we built our training set. If the video contained bleeding and/or lesion/abnormality events, we did not extract from it any false positive samples. We did this us a precaution (and also as simplification of the training set creation), so that our false positive part of the training set did not contain any images from the neighbourhood of the bleeding or lesion events. The colour distribution of these might be altered by the bleeding (altered blood) or contain hard to spot abnormal "redness", which is characteristic for video regions located in the vicinity of lesions. Therefore, in the future it will be possible to retrain the classifiers so that our training set also contained those frames, from the videos containing abnormalities, that we consider the
most definite false positives.

Table 5.2: Performance of SBI vs our method. B, L and FP denote Bleeding, Lesion and False Positive respectively, as number of events and number of frames

<table>
<thead>
<tr>
<th>vid. no.</th>
<th>SBI</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>16</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>20</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>22</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>23</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>24</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

continued on next page
### Table 5.2: continued

<table>
<thead>
<tr>
<th>vid. no.</th>
<th>SBI</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>L</td>
</tr>
<tr>
<td>25</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>26</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>27</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>28</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>31</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>32</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>33</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>34</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>35</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>36</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>37</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>38</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>39</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>40</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>41</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>42</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>43</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>44</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>45</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>46</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>47</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>48</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>49</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>51</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*continued on next page*
### Table 5.2: continued

<table>
<thead>
<tr>
<th>vid. no.</th>
<th>SBI</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>52</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>53</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>54</td>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td>55</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>56</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>57</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>58</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>59</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>60</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>61</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>62</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>63</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>64</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>65</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>66</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>67</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>68</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>69</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>70</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>71</td>
<td>-</td>
<td>16</td>
</tr>
<tr>
<td>72</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>73</td>
<td>-</td>
<td>51</td>
</tr>
<tr>
<td>74</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>75</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>76</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>77</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>78</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*continued on next page*
Table 5.2: continued

<table>
<thead>
<tr>
<th>vid. no.</th>
<th>SBI</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>L</td>
</tr>
<tr>
<td>79</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>80</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>81</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>82</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>83</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>84</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sum</td>
<td>18</td>
<td>136</td>
</tr>
</tbody>
</table>

5.6 Prototype

The research described in this chapter resulted in a prototype of the bleeding detection software. Screenshots of the prototype system can be seen in Figures 5.17 and 5.18 and is due to be deployed late in 2007 for clinical trials in the Gastroenterology Unit of Norfolk and Norwich University Hospital (NNUH), Norwich, UK. The software provides a clinician with a graphical user interface allowing him to execute the bleeding detection algorithm and view the results. As a result of the discussions with the specialist gastroenterologist from NNUH, the emphasis in the design of this prototype was on the user-friendly features that would allow the clinician to navigate easily between different events and frames and consequently shorten the time of the diagnosis.

The SBI function of the Rapid Reader software, marks the detected frames along the video timeline (see Figure 5.19). Then the user can either scroll through the events by clicking certain video control buttons (displayed in red) or click on the bleeding points along the video timeline. It does not make a distinction between bleeding and lesion.
In our software prototype, we decided to create separate screens designated to view the results of running the algorithm. The first screen (see Figure 5.17) displays the detected frames sorted according to the likelihood estimate of frame containing blood - we used here feature $F_2$ defined in Section 5.5. This functionality can take the clinician straight to the place in the video that we consider the most suspicious. If the detected frames do not fit into one page they can be scrolled through page by page.

**Figure 5.17:** Graphic User Interface displaying the results of the blood detection ordered according to the likelihood of blood

In the second screen (see Figure 5.18), the frames are stitched into events and the clinician can go through them either in the time line order or according to the bleeding likelihood estimate order obtained using feature $F_2$ (we take the maximum $F_2$ from among the frames contained in the event and assign it to the event for sorting). The event screen contains the first, last frames of the event as well as the frame with the
maximum $F_2$ from among the frames contained in the event. The clinician can also decide to scroll through the bleeding events first followed by the lesion/abnormality events.

Such a design of the Graphic User Interface can cope with the increased sensitivity of our method i.e. it is particularly useful when a clinician has to scroll through a large number of events.

![Graphic User Interface displaying one of the bleeding events in the video](image)

**Figure 5.18:** Graphic User Interface displaying one of the bleeding events in the video

### 5.7 Conclusions

The results confirm that the bleeding detection method presented in this chapter successfully challenges the state-of-art commercial software. It offers increased sensitivity, without a loss of precision. It detected bleeding in all ten videos containing it as
opposed to seven videos in which SBI was successful. In all bleeding cases more events and frames were detected using our method.

The area in which our method offers an even more significant improvement over SBI is the lesion/abnormality detection in which our method produces more than ten times more detections (calculated on events or frames basis). This has been achieved thanks to the adaptive feature of the pixel classifier, which allows detection of more subtle image features on the changing background as well as advanced region classification, which involves not only the region detected by the pixel classifier, but also its neighbourhood.

The increased sensitivity of the method produces more detections, which can be viewed using the user-friendly software prototype that allows frame and event sorting according to the bleeding likelihood estimate of the detected frame, and the distinction for two classes: bleeding and lesion/abnormality.
It is worth remembering that all these results were achieved using the training set, which although contained hundreds of bleeding images, was produced from only ten full-length videos obtained from the NNUH that contained bleeding and other twenty short sequences downloaded from the Given Imaging website. We strongly believe that the results could be better, had a more representative and significantly larger training set been used. In particular the larger training set of lesions/abnormalities would be useful as this class is visually more variable than the active bleeding.

In the next chapter, we will summarise the findings contained in this thesis. We will also give some details about the possible directions of future work in the field of Wireless Capsule Endoscopy video analysis.
Chapter 6

Conclusions

In this chapter the most important contributions presented in this thesis are summarised. We also place this work in a wider context of WCE image analysis and suggest possible directions for future work.

6.1 Thesis contributions

1. We were the first research group to propose a method for automatic segmentation of the Wireless Capsule Endoscopy video into meaningful anatomical regions.

We have investigated a range of Computer Vision and Image Processing techniques which allowed us to achieve clinically useful segmentation results. These techniques can be divided into feature extraction, image classification and video segmentation stages.

feature extraction - In this stage, we used well studied Hue-Saturation-Intensity (HSI) colour features and 3-D Local Binary Pattern texture/colour features. The features we used were adjusted to our needs i.e. HSI histograms were reduced to the 2-D HS histograms. Moreover, the histograms were also equalised to the range of colours present in the WCE videos covering only around a fifth of the
Hue-Saturation disk.

We also proposed a novel method of extracting these colour and texture features using fixed pre-segmented sub-image blocks, which contain plain (not occluded) tissue only. This feature extraction technique was demonstrated to give some improvement over the feature extraction method taking into account only the entire images.

We proposed a novel method of extracting motion features from the WCE sequence. This method follows the well studied Adaptive Rood Search Pattern, which we used to produce the grid of motion vectors. From the set of motion vectors, we calculated the set of six features, which (when calculated from the certain number of consecutive frames) are further transformed using the Discrete Fourier Transform to provide a final feature vector used in the next classification stage.

Compression was necessary due to the large number of features and consequently different compression methods were used. These included Discrete Cosine Transform (DCT) and Principal Component Analysis (PCA).

*image classification* - This stage involved classification of image features extracted in the previous stage into certain anatomical classes such as: entrance, stomach, small intestine and colon. Here, we investigated different classification schemes which included various linear and non-linear classifiers: Multivariate Gaussian, k Nearest Neighbour and Support Vector Classifier (SVC). We demonstrated that in our task the state-of-art non-linear radial basis SVC yielded the best classification accuracy.

*video segmentation* - Here, we showed how the the sequence of classified features can be used to perform the actual video segmentation, i.e. label the transition points between anatomical organs. In order to achieve this, we investi-
CHAPTER 6. CONCLUSIONS

1. Investigated different segmentation methods which included naïve search, sliding window method and Hidden Markov Model (HMM). The results were obtained using relevant cross-validation. The evaluation included a test of statistical significance (Wilcoxon signed rank test).

2. We proposed a novel algorithm for bleeding detection in Wireless Capsule Endoscopy videos.

Here, the aim was to automatically find those frames in the video that contain bleeding or some related lesions. Our system consists of two parts: adaptive histogram based pixel classifier and the bleeding candidate region classifier.

   adaptive histogram based pixel classifier - We built an adaptive model of blood and non-blood colour distributions. The model was implemented using 3-D HSI histograms. The histogram implementation of the model allows it to be easily updated, according to the changing colour contents of the video. The image pixels whose HSI colour values fall above certain pre-calculated thresholds are considered suspicious and are used in a region growing process to form a candidate bleeding region.

   candidate bleeding region classification - This stage of the algorithm involves extraction of certain low level features (3-D HSI and LBP histograms) from the candidate blood regions and their surrounding neighbourhoods extracted using the morphological dilation operation. Moreover, the candidate region is tested whether it contains specular reflectances that can suggest a presence of the air bubble, which is the indication of possible false detection. Extracted HSI and LBP features are then used to classify the candidate blood region into bleeding, lesion/abnormality and non-blood classes. The classifier used to perform this task was the well known non-linear radial basis Support Vector Classifier. We showed that including the neighbourhood information of the suspicious region into the
feature set significantly improves the classification results. In an extensive comparative experiment, we demonstrated that our method performed better than the state-of-art commercial software. In particular, we showed that a significantly higher number of lesions/abnormalities can be extracted using our method.

6.2 Current state-of-art and future work

In this thesis, we described a number of computer vision algorithms developed by the author as well as presented in the literature by other researchers over the period of last two years. The picture that emerges from this description is very positive as far as the future of this field is concerned. The number of papers published in the field increases fast and there is still plenty of valuable research to be conducted.

Of all the different areas described, arguably the most thoroughly studied was the area of topographic segmentation. The algorithms involved here were tested on real time videos, proofed to perform well and were shown to reduce the video analysis time. In their current form, they could be implemented into real viewing software applications. Indeed, the Endoscopy Unit in Norfolk and Norwich University Hospital (N&NUH) has been equipped with the topographic segmentation software prototype described in Chapter 4. Also in Portugal, three gastroenterology departments (two in public and one in a private hospital) have routine working versions of the CapView 1.0 Annotation Software (CapView, n.d.), which incorporates topographic video segmentation functionality. Regarding future research in this field, the main areas of focused research will involve finding new features, which would allow better discrimination between different tissue types. In particular interesting new prospects might be opened by the use of context features such as capsule location in 3-D, pH factor, temperature etc.

Another major area of capsule research is bleeding detection. Here, a large number of clinical studies have been published, assessing the performance of Given Sus-
CHAPTER 6. CONCLUSIONS

pected Blood Indicator. The authors report different sensitivity and specificity figures, although, they generally agree that the current performance of SBI is insufficient, does not reduce the video viewing time and must be improved. The work on bleeding detection described in Chapter 5 is very promising. It is the only study known to us, which contains a comprehensive comparison of the algorithm performance with the SBI carried out using the large set of full length videos. Moreover, this work demonstrated that it is possible to detect a large number of other abnormalities using this method and in this respect it constitutes a prelude to the solution for detection of any abnormality.

As to detection of other pathological events, the computer vision research is still in its infancy. Neither Given nor Olympus offer any automatic tools capable of detecting abnormalities. The study presented in Section 3.1.2 claims achieving 100% accuracy on detection of abnormalities. However, the size of training and test sets (33 abnormal and 38 normal images) used in this study is insufficient to draw conclusions as to whether the system can be used in a working application. According to us, this is the most challenging field of research in capsule endoscopy image processing since detecting pathologies is the ultimate goal of reliable automatic tools. From the work presented in (Coimbra, Campos and Cunha, 2006a), we can see that this task is very difficult since so far the features (in this particular study MPEG-7 visual descriptors) have not the sufficient discriminant power. The above two studies involved extraction of image features from the entire images. In contrast, the results of our bleeding/abnormality detection experiment suggest that an adaptive framework incorporating both pixel and region/neighborhood based classifier can detect a large number of "reddish" colour related abnormalities. Therefore, it can be expected that the systems of abnormality detection might evolve from this type of framework.

Another successful field of computer vision research is adapting the video play rate to the local video contents. Given Imaging Rapid Reader 4 provides the first automatic tool - Quick View capable of shortening the viewing time. We could not find any clinical
studies assessing the usefulness of this tool. However, given the previously mentioned difficulties regarding detection of single pathologies, this seems to be a reasonable way forward. Moreover, the authors of the publication presented in Section 3.1.5 claim their Quick View-like algorithm achieves reduction of video assessment time to 30 minutes, which is an impressive achievement.

Other algorithms presented in Sections 3.1.1 and 3.1.4 attempt to detect either relevant frames, which should be kept for future viewing by the clinician or irrelevant frames, which may be discarded. Detecting both types could reduce video viewing time - in a particular algorithm, attempting to detect intestinal fluids (see Section 3.1.1), the authors report the mean reduction in a number of frames to be viewed by 23%. Again, this is a very promising result and given what was said in the previous paragraphs about the difficulties of detecting single image pathologies, we strongly anticipate the growth of the number of algorithms, building on the same idea - instead of detecting particular pathologies, we would rather detect and discard irrelevant parts of the video (e.g. intestinal fluids) and again detect and focus attention of the clinician on particularly important frames (e.g. intestinal contractions).

There are still "uncharted territories" in the WCE image processing. All WCE computer vision research so far was focused on the early capsules designated for investigation of the small intestine. The more recent technologies enabling more detailed investigation of the oesophagus and the colon were not investigated in this context. In particular, the latter is expected to attract the significant interest from the Image Processing community.

There are a number of difficulties, that WCE computer vision researchers face. The first problem results from the nature of the capsule video data and is particularly troublesome as far as event detection such as bleeding or abnormality is concerned. The problem is that although each video exam consists of around 50,000 images, these images include very few relevant abnormal events. Thus, it is difficult to build sufficiently
CHAPTER 6. CONCLUSIONS

general models from the large, one might think, set of say around 100 exams. The more specific the abnormality is the more serious this problem becomes. We have come across this problem when building blood colour distributions for bleeding detection. We anticipate, however, that this problem will be even more difficult with regard to describing other pathologies since they are even rarer in the capsule videos than blood and take significantly larger number of forms. Having said that, it is not difficult to explain why abnormality detection, which is the ultimate goal of computer vision in WCE image analysis, have not achieved excellent results in the first two years of capsule video research. This is even more apparent when we look at other areas of WCE computer vision research (such as topographic segmentation, intestinal fluid and contraction detection), where thanks to abundance of relevant data in each video, building general models was possible and consequently resulted in the significant progress in these areas.

The second problem, which will become more apparent in the forthcoming future is the lack of public database of annotated WCE videos, which could be used for testing and comparing the performance of different algorithms as the number of systems addressing similar problems will increase. Building such a database is not an easy task, since the amount of clinicians work annotating hundreds of videos with respect to many types of events (not necessarily relevant to clinicians! e.g. intestinal fluids) is enormous, not to mention the size of the data and resulting server requirements. Moreover, the problem might become even more difficult when the amount of data significantly increases due to unavoidable increases in capsule image acquisition rate (now only 2 frames per second) and image resolution (now $256 \times 256$).

Having said all that, we have 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. With this field clearly maturing and the steady increase in the clinical usage of WCE, we predict a very bright future for
clinical and computer research in this topic.
Appendix A

Wilcoxon signed rank test results of the topographic video segmentation

In this appendix, we describe a statistical test known as the Wilcoxon signed rank test, and how it is applied to confirm the statistical significance of the WCE topographic video segmentation results that were given in Chapter 4 (Tables 4.9 and 4.10). This is followed by the set of figures illustrating the test results for different pairs of video segmentation methods. A discussion of these results can be found in Section 4.8.1.

A.1 Wilcoxon signed rank test

The sign test and the Wilcoxon signed rank tests are the two nonparametric statistical tests concerned with the median of the one sample or paired-sample problems (Gibbons, 1985). In case of the paired-sample problem, which we are dealing with in this work, the ordinary sign test utilises only the signs of the differences between paired samples and ignores their magnitudes. The Wilcoxon signed rank test is an alternative test of median which is affected by both the signs and magnitudes of these differences. In our work, these magnitudes are available, thus the Wilcoxon signed rank test should provide a
better statistical description. The only population assumption is that of symmetry about
the true median or median difference.

The procedure of calculating the signed rank statistic for the paired sample prob-
lem is as follows: The absolute differences between pairs are calculated and ranked.
Any pair of observations giving equal values (zero difference) is ignored. Where ties
occur, the midrank procedure is used i.e. the average of the rank the tied observa-
tions would have if they were not tied is calculated and assigned to them. For exam-
ple, the ordered observations 1, 12, 13, 14, 16, 16, 16, 27 would be assigned ranks
1, 2, 3.5, 3.5, 5, 7, 7, 9. Next, the sums of the ranks with the positive and negative
signs $T_+$ and $T_-$ are calculated. For the one-sided hypothesis test with the null hypoth-
thesis $H_{null} : M = 0$ and the alternative hypothesis $A : M \neq 0$ the P-value is twice the
right tail probability for the larger of $T_+$ and $T_-$. Statistical packages offer calculation
of the exact P-value.

In this work, the P-value of the Wilcoxon signed rank distribution was calculated
using the modified signrank function from Matlab 7.0. The figures in this appendix
illustrate the test of the null hypothesis $H_{null} : M_D = 0$ for the alternative $A_- : M_D < 0$,
where $D = Y - X$ and hence the P-values given in them correspond to right-tail
probabilities of $T_-$ or left-tail of $T_+$. $X$ and $Y$ correspond respectively to the first and
the second method being compared. Thus, we are testing $H_{null}$ with the alternative
hypothesis that the median of paired error differences is smaller than zero (the second
method is better than the first).

### A.2 Wilcoxon signed rank test results of topographic video segmentation

In this section, after introducing the figure notation scheme, we provide a series of
Wilcoxon signed rank test results divided into six groups represented by the following
APPENDIX A. SIGNED RANK TEST RESULTS

subsections.

Notes on figure notation

Each of the remaining figures in this appendix consists of six sub-figures. Those on the left illustrate the histograms of errors of the two methods being compared. Those on the right show the number of videos where the error of the second method (second row of the legend) has decreased, stayed the same or increased in comparison with the first method (first row of the legend). The entry of each method in the legend consists of the segmentation method name e.g. naïve, followed by the classifier name e.g. SVC and followed by the feature vector name e.g. HS. They can be optionally followed by some additional information about the method. For the naïve classifier, it can be either one pass or multiple pass and for sliding window method, these are the three sizes (entrance/stomach, stomach/intestine and intestine/colon) of the sliding windows. In addition, each of the three rows in any figure contains the medians of the two methods and the P-value calculated as described in Section A.1.

A.2.1 Naïve (1 pass) versus Naïve (multiple passes)

In this section we compare video segmentation results for the single pass versus the multiple passes of the naïve algorithm.
Figure A.1: Signed rank test results for naïve (1 pass) versus naïve (multiple passes) methods using LBP 3D feature vectors.
Figure A.2: Signed rank test results for naïve (1 pass) versus naïve (multiple passes) methods using HS feature vectors.
Figure A.3: Signed rank test results for naïve (1 pass) versus naïve (multiple passes) methods using DFT feature vectors.
A.2.2 Naïve versus Sliding Window

In this section we compare the video segmentation results obtained for the multiple pass naïve algorithm versus the sliding window method.

Figure A.4: Signed rank test results for naïve (multiple passes) versus sliding window (window sizes: 50, 200, 1200) methods using HS feature vectors
Figure A.5: Signed rank test results for naïve (multiple passes) versus sliding window (window sizes: 50, 1200, 3000) methods using HS feature vectors.
APPENDIX A. SIGNED RANK TEST RESULTS

Figure A.6: Signed rank test results for naïve (multiple passes) versus sliding window (window sizes: 50, 200, 1200) methods using LBP 3D feature vectors
**Figure A.7:** Signed rank test results for naïve (multiple passes) versus sliding window (window sizes: 50, 1200, 3000) methods using LBP 3D feature vectors
A.2.3 Sliding Window versus HMM

In this section we compare the video segmentation results obtained for the sliding window method versus the HMM method.

![Figure A.8: Signed rank test results for sliding window (window sizes: 50, 200, 1200) versus HMM methods using LBP 3D feature vectors](image-url)
APPENDIX A. SIGNED RANK TEST RESULTS

Figure A.9: Signed rank test results for sliding window (window sizes: 50, 1200, 3000) versus HMM methods using LBP 3D feature vectors
Figure A.10: Signed rank test results for sliding window (window sizes: 50, 200, 1200) versus HMM methods using HS feature vectors.
Figure A.11: Signed rank test results for sliding window (window sizes: 50, 1200, 3000) versus HMM methods using HS feature vectors.
APPENDIX A. SIGNED RANK TEST RESULTS

Figure A.12: Signed rank test results for sliding window (window sizes: 50, 200, 1200) versus HMM methods using LBP 3D + HS feature vectors
Figure A.13: Signed rank test results for sliding window (window sizes: 50, 1200, 3000) versus HMM methods using LBP 3D + HS feature vectors
A.2.4 Multivariate Gaussian versus SVC

In this section we compare the video segmentation results obtained for the Multivariate Gaussian classifier versus the SVC. Both classifiers were used together with the HMM video segmentation method.

Figure A.14: Signed rank test results for Multivariate Gaussian versus SVC methods using HMM video segmentation and HS feature vectors
APPENDIX A. SIGNED RANK TEST RESULTS

Figure A.15: Signed rank test results for Multivariate Gaussian versus SVC methods using HMM video segmentation and LBP 3D feature vectors.
APPENDIX A. SIGNED RANK TEST RESULTS

Figure A.16: Signed rank test results for Multivariate Gaussian versus SVC methods using HMM video segmentation and LBP 3D + HS feature vectors
Figure A.17: Signed rank test results for Multivariate Gaussian versus SVC methods using HMM video segmentation and DFT feature vectors.
Figure A.18: Signed rank test results for Multivariate Gaussian versus SVC methods using HMM video segmentation and LBP 3D + HS + DFT feature vectors.
APPENDIX A. SIGNED RANK TEST RESULTS

A.2.5 Entire image features versus entire image plus sub-image region features

In this section we compare the WCE video segmentation results obtained for the features extracted from the entire images versus the features extracted from the entire images and Sub-image region regions. Both methods were used together with the SVC classifier, the radial basis kernel and the HMM video segmentation method.

Figure A.19: Signed rank test results for HS feature vectors extracted from entire images versus HS feature vectors extracted from entire images and SubIRs using SVC and HMM video segmentation.
Figure A.20: Signed rank test results for LBP 3D feature vectors extracted from entire images versus the same feature vectors extracted from entire images and SubIRs using SVC and HMM video segmentation.
APPENDIX A. SIGNED RANK TEST RESULTS

Figure A.21: Signed rank test results for LBP 3D + HS feature vectors extracted from entire images versus the same feature vectors extracted from entire images and SubIRs using SVC and HMM video segmentation.
Figure A.22: Signed rank test results for LBP 3D + HS + DFT feature vectors extracted from entire images versus the same feature vectors extracted from entire images and SubIRs using SVC and HMM video segmentation.
A.2.6 Different features

In this section we are comparing video segmentation results obtained for different types of features such as HS histograms, LBP 3D histograms, DFT features as well as the combinations of these three types of features. All the results were obtained for the HMM video segmentation method with the radial basis kernel SVC classifier.

Figure A.23: Signed rank test results for HS feature vectors versus LBP 3D feature vectors using SVC classifier and HMM video segmentation
**APPENDIX A. SIGNED RANK TEST RESULTS**

Figure A.24: Signed rank test results for DFT feature vectors versus HS feature vectors using SVC classifier and HMM video segmentation.
Figure A.25: Signed rank test results for HS feature vectors versus LBP 3D + HS feature vectors using SVC classifier and HMM video segmentation.
APPENDIX A. SIGNED RANK TEST RESULTS

Figure A.26: Signed rank test results for LBP 3D feature vectors versus LBP 3D + HS feature vectors using SVC classifier and HMM video segmentation.
Figure A.27: Signed rank test results for LBP 3D + HS feature vectors versus LBP 3D + HS + DFT feature vectors using SVC classifier and HMM video segmentation.
References


REFERENCES


CapView (n.d.). Portugal.  
URL: http://www.capview.org/


REFERENCES

Coimbra, M., Campos, P. and Cunha, J. P. S. (2005), Extracting clinical information from endoscopic capsules exams using MPEG-7 visual descriptors, in ‘2nd European Workshop on the Integration of Knowledge Semantic and Digital Media Technologies’, IEE.


REFERENCES


Given Imaging Ltd (n.d.). Israel. URL: http://www.givenimaging.com


Innovative imaging probes for endoscopy (2005), European Innovations (4).

Intracorporeal Videoprobe Project (n.d.).

URL: http://ivp.ims-chips.de/Results/IVP.html


Lin, H., Lin, C. and Weng, R. (2003), A note on Platt’s probabilistic outputs for support vector machines, Technical report, Department of Computer Science and Information Engineering, National Taiwan University. [URL: citeseer.ist.psu.edu/lin03note.html](http://citeseer.ist.psu.edu/lin03note.html)


REFERENCES


REFERENCES


REFERENCES


van Rijsbergen, C. J. (1979), Information Retrieval, 2 edn, Butterworth and Co Ltd.


REFERENCES


