DeepEIR: A Holistic Medical Multimedia System for Gastrointestinal Tract Disease Detection and Localization

Konstantin Pogorelov

17.07.2019

© Konstantin Pogorelov, 2019

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 2195

ISSN 1501-7710

All rights reserved. No part of this publication may be
reproduced or transmitted, in any form or by any means, without permission.

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Abstract

Advanced and automated medical systems have been in the research focus for a long time. Together with the rapid development of sensing devices, the modern information analysis methods allow the new wave of computer-assisted systems to improve health care, quality of life, and patient survival rate. Together with the traditional computer vision and medical imaging, core competencies of the multimedia community such as integration and analysis of data from several sources, real-time processing and the assessment of usefulness for end-users play an essential role for the successful improvement of health care systems addressing challenges and open problems in the field of medicine.

Our work explores different fields in multimedia research, starting from collection and annotation of multimedia data through automatic analysis of content and efficient processing of workloads to visualization and results representation. We have researched and developed a holistic medical multimedia system addressing a use case with an important medical and societal impact. We target lesions and findings detection and localization in the gastrointestinal (GI) tract of the human body in order to be able to support medical experts in their daily routine work. The early and precise detection of abnormalities in the GI tract greatly increases the chance of successful treatment if the initial observation of disease indicators occurs before the patient notices any symptoms, it is a non-trivial task that can be, however, efficiently automated.

We investigated the GI tract visual analysis from a multimedia research point of view via several steps of research and development. First, we looked into the problem of medical data acquisition. We collected, annotated, and published several datasets and data annotation tools as open source. Then, we designed and developed a set of lesion and findings detection and localization approaches based on hand-crafted methods as well as on global-, local- and deep-feature-based methods, which serves as the algorithmic basis of our system. Next, we created a holistic medical multimedia system called DeepEIR. We researched and developed different subsystems for our DeepEIR system, namely (i) the data exploration and annotation subsystem, which makes it possible to collect and annotate data and transfer knowledge from medical experts into our system; (ii) the detection and localization subsystem, which perform medical data analysis in order to detect and localize lesions and findings; and (iii) the visualization and results representation subsystem that provides the information to medical personnel.

Furthermore, the focus of the DeepEIR system lies on the accurate and time-efficient processing of multimedia data. We investigated, therefore, parallel and distributed processing, GPU-based acceleration and different classification and segmentation approaches that are evaluated and compared with state-of-the-art methods, algorithms, and systems.

We demonstrated that the DeepEIR system could outperform state-of-the-art approaches in both processing speed and detection accuracy reaching processing speeds above 300 frames per second, a frame-wise detection accuracy above 95% and pixel-wise localization accuracy above 90%. With our results good enough for the clinical trials and successful demonstration of full-scale prototypes of DeepEIR system, we were able to attract several hospitals for tight collaborations, and the DeepEIR system is being prepared for a broad testing and using under clinical conditions within our collaborating hospitals.

Acknowledgements

First, I would like to thank my three official PhD supervisors: Pål Halvorsen, Carsten Griwodz and Michael Riegler. I would like to especially thank: Pål for his supervision, useful advice and support. Carsten for his critical yet guiding discussions and feedback. Michael for being a research partner during our well-established work together.

I would like to especially thank my current research supervisor Johannes Langguth for the provided possibility to finish my PhD thesis writing while working on our new research project.

I would also like to give many thanks to all my current and former colleagues in Simula Research Laboratory. We were working, talking and having fun together. Thank you, my dear Vamsi, Jonas, Preben, Håkon, Iffat, Andreas, David, Minoo, Olga, Ragnhild, Kjetil, Lilian, Robin, Vajira, Debesh and Steven.

Big thank you, people of Norway, for preserving the nature - the most valuable Earth's resource.

And finally, but the most important, I would like to say "Thank you so much!" to my parents for their infinite support of my curiosity and interest in science and tech. Thank you, Liudmila and Vladimir!

I esteem myself happy to have as great an ally as you in my search for truth.

Galileo Galilei

Contents

I Overview 1
1 Introduction 3
  1.1 Background and Motivation 4
  1.2 Problem Statement 7
  1.3 Scope and Limitations 9
  1.4 Research Methods 10
  1.5 Contributions 11
  1.6 This thesis author’s independent contributions 15
  1.7 Outline 17
2 Medical Multimedia Systems 19
  2.1 Gastrointestinal Tract Case Study 19
    2.1.1 Endoscopic Devices 21
      2.1.1.1 Traditional Endoscopes 21
      2.1.1.2 Wireless Video Capsular Endoscopes 22
    2.1.2 Medical Data 24
  2.2 Medical Image Analysis 25
    2.2.1 Challenges of Automatic Diseases Detection 25
    2.2.2 State of the Art in GI Tract Lesion Detection 26
    2.2.3 Basic EIR System: The Proof-of-Concept 29
  2.3 Summary 29
3 The DeepEIR System 31
  3.1 Data Collection 32
    3.1.1 Privacy, Legal and Ethics Issues 33
    3.1.2 Sources of the Data 33
    3.1.3 Created Datasets and Reproducibility 34
      3.1.3.1 Kvasir 34
      3.1.3.2 Nerthus 37
      3.1.3.3 Medico Task 38
      3.1.3.4 Further Dataset Development 41
      3.1.3.5 Application of the Datasets 43
  3.2 Data Exploration, Annotation and Visualization Subsystem 43
    3.2.1 Hyperbolic-Tree-Based Visualization and Clustering 44
    3.2.2 Cluster-Based Visualization and Annotation (ClusterTag) 46
  3.3 Detection Subsystem 48
    3.3.1 Single-Class Global-Feature-based Detection 49

3.3.2 Multi-Class Global-Feature-based Detection 50
3.3.3 Deep-Learning-based Detection 51
3.3.4 Deep-Feature-based Detection 52
3.4 Localization Subsystem 53
3.4.1 Hand-Crafted Local-Feature-based Position Finder 53
3.4.2 Deep-Learning-based Region Localizers 54
3.4.3 Deep-Feature-based Region Localization 56
3.4.4 GAN-based Segmentation, Localization and Detection 56
3.5 Visualization and Results Representation Subsystem 59
3.5.1 Online Global-Feature-Based Visual Similarity Search Tool 59
3.5.2 Visualization Module for Polyp Detection and Spotting 59
3.5.3 Visualization Module for Lesions Detection and Localization 60
3.6 System Evaluation 62
3.6.1 Annotation Subsystem 63
3.6.2 Detection and Localization Subsystems 64
3.6.2.1 Evaluation Metrics 64
3.6.2.2 Polyps 64
3.6.2.3 Angiectasia 70
3.6.2.4 Multi-Class Detection 74
3.6.3 Detection Subsystem Processing Speed Optimization 78
3.6.3.1 Heterogeneous Architecture 80
3.6.3.2 Processing Speed Evaluation 81
3.6.3.3 Distributed Heterogeneous Architecture 84
3.6.4 System Extensibility Test 87
3.6.4.1 Bladder Cancer Cells Detection and Localization 88
3.6.4.2 Spermatozoon Localization and Segmentation 90
3.7 Summary 93
4 Conclusion 97
4.1 Summary and Contributions 97
4.2 Future Work 101
4.3 Final Remarks 102
5 Papers and Author’s Contributions 103
5.1 Paper I: LIRE - Open Source Visual Information Retrieval 103
5.2 Paper II: OpenSea - Open Search Based Classification Tool 104
5.3 Paper III: Explorative Hyperbolic-Tree-Based Clustering Tool for Unsupervised Knowledge Discovery 104
5.4 Paper IV: ClusterTag: Interactive Visualization, Clustering and Tagging Tool for Big Image Collections 105
5.5 Paper V: EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal Endoscopies 106
5.6 Paper VI: From Annotation to Computer-Aided Diagnosis: Detailed Evaluation of a Medical Multimedia System 106
5.7 Paper VII: Multimedia and Medicine: Teammates for Better Disease Detection and Survival 107

- 5.8 Paper VIII: A Holistic Multimedia System for Gastrointestinal Tract Disease Detection . . . 108
- 5.9 Paper IX: GPU-accelerated Real-time Gastrointestinal Diseases Detection . . 109
- 5.10 Paper X: Efficient Processing of Videos in a Multi-Auditory Environment Using Device Lending of GPUs . . . 110
- 5.11 Paper XI: Efficient disease detection in gastrointestinal videos - global features versus neural networks . . . 110
- 5.12 Paper XII: Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection . . . 111
- 5.13 Paper XIII: Nerthus: A Bowel Preparation Quality Video Dataset . . . 112
- 5.14 Paper XIV: Deep Learning and Handcrafted Feature Based Approaches for Automatic Detection of Angiectasia . . . 113
- 5.15 Paper XV: Deep Learning and Hand-crafted Feature Based Approaches for Polyp Detection in Medical Videos . . . 114

II Research Papers . . . 129

I LIRE - Open Source Visual Information Retrieval . . . 131

II OpenSea - Open Search Based Classification Tool . . . 137

III Explorative Hyperbolic-Tree-Based Clustering Tool for Unsupervised Knowledge Discovery . . . 145

IV ClusterTag: Interactive Visualization, Clustering and Tagging Tool for Big Image Collections . . . 151

V EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal Endoscopies . . . 159

VI From Annotation to Computer-Aided Diagnosis: Detailed Evaluation of a Medical Multimedia System . . . 167

VII Multimedia and Medicine: Teammates for Better Disease Detection and Survival . 195

VIII A Holistic Multimedia System for Gastrointestinal Tract Disease Detection 207

IX GPU-accelerated Real-time Gastrointestinal Diseases Detection . . . 221

X Efficient Processing of Videos in a Multi-Auditory Environment Using Device Lending of GPUs . . . 229

XI Efficient disease detection in gastrointestinal videos - global features versus neural networks . . . 235

XII Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection . . . 271

XIII Nerthus: A Bowel Preparation Quality Video Dataset . . . 279

XIV Deep Learning and Handcrafted Feature Based Approaches for Automatic Detection of Angiectasia . . . 287

XV Deep Learning and Hand-crafted Feature Based Approaches for Polyp Detection in Medical Videos . . . 293

List of Figures

1.1 An overview of the human GI tract (hdfootagestock.com). 4

1.2 An inconclusive list of diseases that can be observed and diagnosed in GI tract [95]. These are the real images recorded from endoscopic equipment during routine examinations. Green box shows the status a colonoscope device. 5

1.3 Colonoscopy is the endoscopic examination (a) of the large bowel and the distal part of the small bowel with a special type endoscope called coloscope (b) [116]. 6

1.4 Capsule endoscopy is a non-invasive procedure used to record internal images of the GI tract using a small swallowed VCE device equipped with a camera, a battery and a transmitting or recording module [116]. 7

1.5 This diagram depicts the contributions for each of the in part II attached papers to the, for this thesis defined, objectives. 12

2.1 The internal components of wireless video capsule endoscope 23

3.1 A complete overview of the DeepEIR system. The system consists of data acquisition, preparation and annotation, automatic analysis and visualization subsystems. 32

3.2 Sample images of the GI tract lesions included in the Kvasir dataset. 36

3.3 Sample images of the GI tract landmarks included in the Kvasir dataset. 37

3.4 Sample images for each bowel preparation ("cleanliness") score according to BBPS. 38

3.5 The example images depicting different amount of stool masses in the colon. . . 40

3.6 Images depicting various instruments including manipulating devices (a) and (b), and endoscope itself captured via retroflex action (c) and (d). 41

3.7 Images depicting auxiliary image classes: (a) blurry frames without any recognizable content, and (b) out of the patient images. 42

3.8 Images depicting various classes will be added to our open datasets in the near future 42

3.9 Hyper-tree based visualization, clustering and annotation system. 45

3.10 Structure of the visualization and user interface engine of the presented ClusterTag application. A number of caching and intermediate data processing routines are used to make it possible to perform real-time visualization and interaction with huge image collections. 46

3.11 Examples of visual representations of an image collection containing $36,476$ unsorted medical images generated by the ClusterTag application. The initial view of the loaded collection shows all the images in one big cluster. After the clustering, using the JCD and Tamura global image features, the software generates a number of dense clusters representing visually similar images in the same clusters. 48

3.12 Detailed steps for the multi-class global-feature-based detection implementation 50

3.13 Multi-class deep-learning-based detection pipeline 52

3.14 DCNN concepts- and deep-features-based detection pipeline 53

3.15 Detailed steps of the hand-crafted local-feature-based localization algorithm implementation 54

3.16 Example frames showing polyp and its body ground truth area. This is an example of polyps localization task complexity. Polyp body has the same color, texture properties and light flares as surrounding normal mucosa 56

3.17 GAN-based segmentation and localization pipeline 57

3.18 Examples of the different auxiliary information fields integrated into recorded frame: a colonoscope navigation localizer (a), a captured still frame (b) and a patient-related information (c). Images taken from CVC-968 [23] and Kvasir [95]. 58

3.19 GAN-based detection-via-localization pipeline 58

3.20 Online global-feature-based visual similarity search tool usage examples. The image in the center is the query image. The first six results of four queries based on four three global and one local features are shown around the query image. . 60

3.21 The visualization module for real-time polyp detection and spotting build upon our global-feature-based detection and hand-crafted local-feature-based polyp position finder approaches. It is able to process both recorded and live Full HD video stream from traditional colonoscope, highlight frames containing polyps and mark the recognized polyp location with a cross mark. The pink surrounding frame shows a positive detection. Plot in the lower part of UI shows the per-frame polyp presence ground truth, polyp detection indicator and TP/FP/FN/TN events recorder. 61

3.22 The visualization module for our deep-feature-based real-time polyp detection approaches. It is able to process Full HD live-captured video stream from traditional colonoscope and highlight frames containing detected lesions. The plot in the lower part of UI show the per-video-frame lesion detection probability. . 61

3.23 Near-to-real-time polyp detection and localization demo build upon our GAN-based detection and localization approach. The software processes recorded Full HD video stream from traditional colonoscope and highlights the exact polyp location in the particular frame. The marking is implemented as as a bounding box rectangle drawing over the source video frame. The achieved processing speed is in between 5 and 10 FPS depending on the used GPU acceleration hardware. 62

3.24 Polyp localization results generated by our first polyp localization and detection approach on the MICCAI 2015 dataset [25]. Light green ellipses depicts the polyp localization ground truth masks. Green and red crosses show the true positive and false positive polyp localization results, respectively. The localization algorithm was tuned to output exact four possible polyp locations per frame. 65

3.25 The example ot the polyp localization mask generated by our GAN-based polyp localization approach. The base polyp localizer generates the pixels-wise probability mask shows the possible localization of the polyp body's pixels. The green ellipse highlights the polyp body for illustration purposes only. The resulting localization mask conforms good with the ground truth. 69

3.26 Example of difficult images in the test dataset: a significant frame blur caused by camera motion (a), a color components shift caused by the temporary signal failure (b) and an out-of-focus frame contains also contamination on the camera lens (c). Images taken from the CVC-12k [23]. 70

3.27 Examples of the detection and in-frame localization of the different polyps in the video frames captured by various vendors' traditional colonoscopy equipment. Green contour depicts the detected polyp and the localized main polyp body area. 71

3.28 Example of an angiectasia lesion marked with a green circle (a), a corresponding ground truth mask (b) and a segmentation mask generated using our GAN-based approach (c). Image taken from the GIANA dataset [22]. 72

3.29 Examples of the detection and in-frame localization of the clearly visible angiectasia areas. 73

3.30 Examples of the detection and in-frame localization of the partially obscured, tiny and hard-to-spot angiectasia areas. 73

3.31 The main processing application consisting of the indexing and classification parts uses the GPU-accelerated image processing subsystem. This subsystem provides feature extraction and image filtering algorithms. The most compute-intensive procedures are executed on a stand-alone CUDA-enabled processing server. The interaction between application and server is done via a GPU CLib shared library, which is responsible for maintaining connections and streaming data to and from the CUDA-server. 80

3.32 GPU-acceleration is used to extract various features from input frames. The figure shows an example of our FCTH feature implementation. The input frame is split into a number of non-overlapping blocks. Each of them is processed separately by two GPU-threads. The main processing steps include color space conversion, size reduction, shape detection and fuzzy logic computations. . . . 81

3.33 The detection performs efficiently and the required frame rate is reached with 12 GB of memory and 16 CPU cores used in parallel on cluster-based computation platform without utilizing heterogeneous architecture. 82

3.34 The improved GPU-enabled heterogeneous algorithm reaches real-time performance (RT line) with 30 frames per second for full HD (1920 × 1080) videos on a desktop PC using only 4 CPU cores and 5 Gb of memory. The maximum frame rate is around 36 FPS using 8 CPU cores. The Java and C implementations cannot reach real-time performance on the used hardware. 83
3.35 The smaller WVGA1 (856 × 480), WVGA2 (712 × 480) and CIF (384 × 288) videos can be processed by the improved GPU-enabled heterogeneous algorithm in real-time using only 1 CPU core. The maximum frame processing rate reaches more than 200 FPS. These results can be improved by putting all feature-related computations on the GPU. 83
3.36 The processing time for the GPU-accelerated algorithm decreases slightly with increasing number of used CPU cores for a single full HD frame. This happens due to the CPU-parallel implementation of feature comparison and search algorithms which are not as compute intensive as feature extraction. The Java and C implementations reach the minimum frame processing time with 4 used CPU cores. The reason is that the used CPU has 4 real cores with hyper-threading feature enabled and it cannot handle CPU-intensive calculations efficiently for all 8 (real plus virtual) cores. 84
3.37 For the smaller frame sizes the GPU-accelerated algorithm results in a processing time far below the real-time margin. The minimum is reached with 5 milliseconds using 8 CPU cores. This is a prove for the high system performance and ability to be extended by additional features or to process several video streams at the same time on a conventional desktop PC. 85
3.38 Pooling of devices attached in the PCIe network in the experimental setup. . . . 85
3.39 System performance evaluation in terms of processing time per frame and maximum performance using 4 different configurations described in table 3.21. Each video stream is a full HD video. 86
3.40 The examples of WLC (a) and BLC (b) frame of our dataset used for the experimental evaluation of the EIR system flexibility and extendability. Images (a) and (b) contain the instrument tip visible in the image top-right corner. Tumor cells clusters are colored by pink color and located in the middle (b), in the middle and top-center (c), and around of the middle (d) of the images. 88
3.41 The examples of the localized clusters of the bladder cancer cells. The green boxes in the images mark the successfully recognized tumors' locations including ones on the side of the field of view (c), bedly visible in the dark areas (a), located on the blood vessels (b) and partially covered by the tissue (d). One tiny group of cells is missed (e, top-center) probably because of bad input image quality caused by strong video encoding. Constantly visible similarly colored not detected objects are the standard instrument tips. 89
3.42 The example images of the spermatozoon localization and segmentation dataset used for the experimental evaluation of the EIR system with the different use-case study. First image (a) depicts the source microscopic image in RGB color space. Three other images (b-d) represent the ground truth masks for the different morphological parts of the spermatozoons shown on the image (a). 91

3.43 The comparison of the ground truth segmentation masks with the output generated segmentation masks of the different morphological parts of the spermatozoons. 92

List of Tables

2.1 Existing endoscopic image and video datasets 25

2.2 A performance comparison of GI findings detection approaches. Not all performance measurements are available for all methods, but including all available information gives an idea about each method's performance. Also there are many done and ongoing research in the field, and this table present a selection of the most representative and recent results 27

3.1 Results of the MICCAI 2015 polyp localization challenge [25]. 64

3.2 Results of the MICCAI polyp detection challenge. The table shows the detection latency in milliseconds and F1 score [25]. 65

3.3 Overview of the datasets used in the experiments. Kvasir and Nerthus are our own public datasets. CVC-968 is a combined dataset consist of CVC-356 and CVC-612 sets. 67

3.4 Validation results of the in-frame pixel-wise polyp areas segmentation (localization) approach evaluated using different combinations of the CVC-356 and CVC-612 sets for training and testing. 67

3.5 Performance of the block-wise polyp localization (LOC) via detection approaches reported per method and used training data. Training and testing are performed using the CVC-968 and CVC-12k datasets, respectively. See Paper XV for the detailed results. 67

3.6 Results for the frame-wise polyp detection approaches, namely multi-class global-feature-based (GFD), deep-learning-based with random tree (RTD) final classifier, GAN-based (GAND) and YOLOv2-based (YOLOD). We used the CVC-12k and Kvasir dataset as independent test sets. Training of all the approaches is performed using the combined CVC-968 dataset consist of CVC-356 and CVC-612 sets. See Paper XV for the detailed results. 68

3.7 This table depicts performance of the in-frame pixel-wise polyp localization (segmentation) approach evaluated using different combinations of the CVC-356 and CVC-612 datasets for training and testing. 69

3.8 This table depicts performance of the block-wise localization via detection approach for the CVC-12K dataset reported for different training data used. . . . 70

3.9 This table depicts performance of the frame-wise polyp detection approach. We used different small training sets and the CVC-12k and Kvasir dataset as independent test sets. 70

3.10 This table depicts ten-fold cross-validation results of the pixel-wise GAN-based angiectasia localization approach (the 95% confidence intervals are reported). See Paper XIV for the detailed results. 72
3.11 This table depicts ten-fold cross-validation results of the angiectasia frame-wise detection using the GAN approach (the 95% confidence intervals are reported). See Paper XIV for the detailed results. 72
3.12 Results for the angiectasia frame-wise detection approaches evaluated with the annotated test set. See Paper XIV for the detailed results. 73
3.13 A confusion matrix for the six-classes detection performance evaluation for the Deep-EIR detection subsystem 74
3.14 Performance evaluation of the six-classes detection for the Deep-EIR detection subsystem 75
3.15 A confusion matrix for the six-classes detection performance evaluation for the multi-class global-feature-based EIR detection subsystem 75
3.16 Performance evaluation of the six classes detection for the multi-class global-feature-based EIR detection subsystem 75
3.17 Performance evaluation of the cross-validation for the Deep-EIR and the multi-class global-feature-based EIR detection subsystems 76
3.18 The per-class-contents of the training and test dataset used for the multi-class detection algorithms evaluation. This dataset was used for the Medico task at MediaEval 2018 contest [100]. 77
3.19 Classification performance evaluation for the detection models, trained using the augmented (A) and size-equalized (E) training sets including ZeroR (ZR), Random (RD) and True (TR) baseline classifiers. Runs #1 corresponds to the non-prioritized classification, while runs #2 - #5 corresponds to the 0.75 to 0.1 classification probability threshold level. 78
3.20 Confusion matrix for the run A1 depicted in table 3.19. The classes are Ulcerative Colitis (A), Esophagitis (B), Normal Z-line (C), Dyed and Lifted Polyps (D), Dyed Resection Margins (E), Out of Patient images (F), Normal Pylorus (G), Stool Inclusions (H), Stool Plenty (I), Blurry Nothing of value (J), Polyps (K), Normal Cecum (L), Colon Clear (M), Retroflex Rectum (N), Retroflex Stomach (O) and Instruments (P). 79
3.21 This table shows the used hardware combinations of the different experiments. GPU 1 to 3 are local GPUs. GPU4 is lend via Device Lending. 87

Part I

Overview

Chapter 1

Introduction

In current modern life, we all are surrounded by a huge amount of data. The dominating one is the multimedia data and, especially, visual data in forms of images and videos. The constant progress in the fields of computer vision, information retrieval and understanding already resulted in a variety of efficient methods that can utilize such the data and produce a broad range of valuable output ranging from face recognition for social networks and security systems to remote sensing application that are able to detect disasters in remote areas using satellite imagery. The estimated size of data in the health care system for the whole world is around 162 exabyte, with an estimated increase of 2.5 exabytes per year [27]. A significant part of this data is producing by the health care system with the increasing speed. The future gigantic scale of medical data [117] comes with several challenges to analyze, store, transmit and utilize it for useful purposes. However, the challenges should be addressed as soon as possible to bring the advantages related to the multimedia data processing to the current healthcare system.

Some of multimedia data challenges in medicine are collecting, understanding and analyzing data, and reusing the medical knowledge. Next, the practical challenges of performance and real-time processing speed come to the front during the implementation of the real systems for live patient examination, communication, or other medical tasks. Even the very modern visual data processing and understanding methods cannot be efficient enough yet because of both under-development and lack of available training data. Another need that comes with a large amount of data is efficient, robust and scalable data processing methods. Because of a large amount of multimedia data in the health care system, parallel processing and elastic heterogeneous resources are important [117] to achieve fast processing of multimedia workloads by being able to process a large amount of data in parallel at the same time.

In this work, we investigate how the new computer vision and machine learning methods can be utilized and improved in order to build a completely automatic diagnostic assisting system that is able to support medical experts in disease detection, live patient examinations and national-wide screening programs. Since the medical field by itself is enormous, we decided to address one area in this field specifically. We decided on the human gastrointestinal (GI) system because it can potentially be affected by many types of diseases that are visually distinguishable. This choice is also supported by the fact that the most common cancer types are located in the GI tract [148]. An accurate automatic medical analysis system will have a high impact on the medical sector, influencing patient survival rates, clinical workflows and costs. In the GI field, medical imaging has created visual representations of the interior of a body with images, videos and corresponding text descriptors made by doctors during routine procedures. This work focuses on investigating efficient analysis and processing of multimedia workloads in the field of GI endoscopy with the goal of creating new methods and a complete prototype of an end-to-end medical multimedia system that will assist doctors during GI tract investigations.

1.1 Background and Motivation

The modern healthcare system has been intensively improved during the last decades, introducing a lot of different modern diagnostic methods. However, there are a lot of unsolved medical and societal challenges still affecting the effectiveness of the health care systems worldwide. In some areas of the human body, such as the gastrointestinal (GI) tract (figure 1.1), the detection of abnormalities and diseases directly improves the chance of successful treatment.

The GI tract diagnosis is important since it is the site of many common diseases (see figure 1.2 for the examples) with high mortality rates. About 2.8 million new luminal GI cancers (esophagus, stomach, colorectal) are detected yearly in the world, and the mortality is about 65% [50]. In addition to these cancers, numerous other chronic diseases affect the human GI tract. The most common ones include gastroesophageal reflux disease, peptic ulcer disease, inflammatory bowel disease, celiac disease and chronic infections. All these diseases have a significant impact on the patients' health-related quality of life [34] and, therefore, gastroenterology is one of the critical and largest medical branches.

For the most severe, colorectal cancer (CRC), which has one of the highest incidences and mortality of the diseases in the GI tract, early detection is essential for a good prognosis and treatment. Minimally invasive endoscopic and surgical treatment is most often curative in early stages (I-II) with a 5-year survival probability of more than 90%. But in advanced stages (III-IV), radiation and/or chemotherapy is often required, and it has a 5-year survival of only 10-30% [30]. Moreover, several studies have shown that large population-based endoscopic screening programs reduce the mortality and incidence of CRC. The current European Union guidelines, therefore, recommend screening for CRC [144]. Several screening methods exist, e.g., fecal immunochemical tests (FITs), sigmoidoscopy screening, computer tomography (CT)

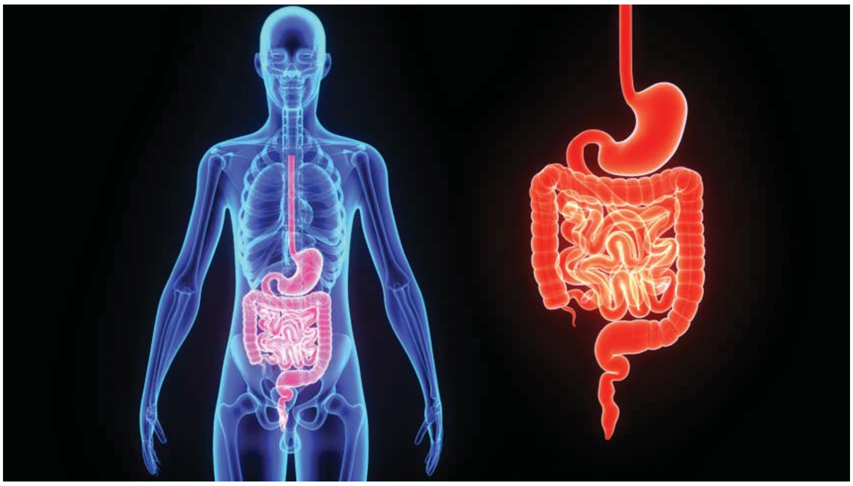
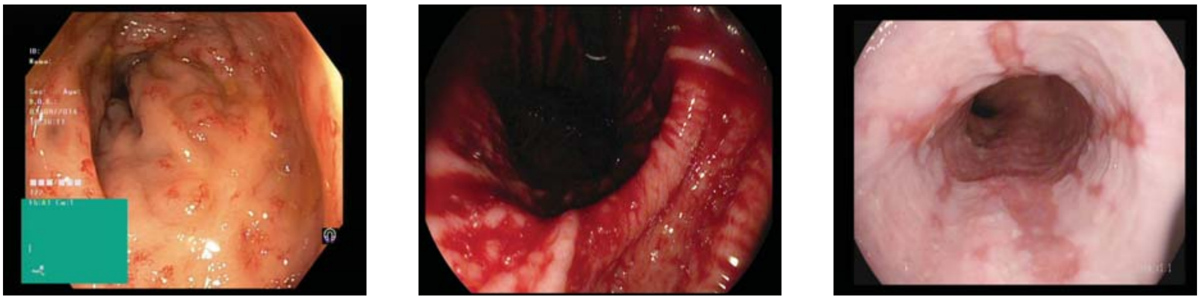
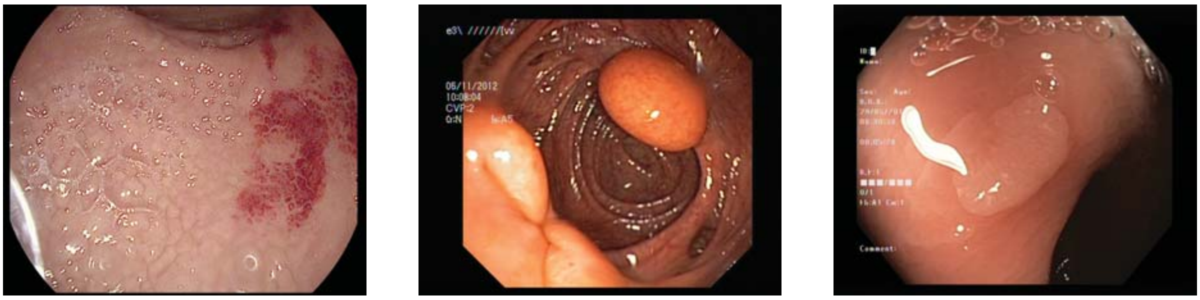


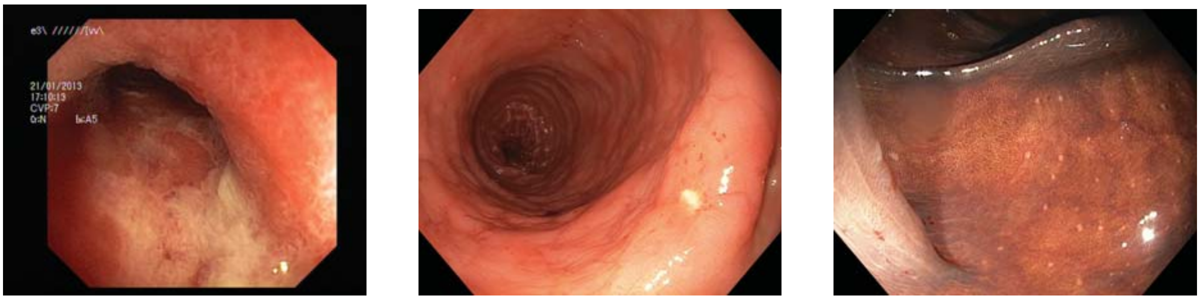
Figure 1.1: An overview of the human GI tract (hdfootagestock.com).



(a) Angiectasia (b) Bleeding (c) Esophagitis



(d) Inflamation (e) Polyp (f) Flat polyp



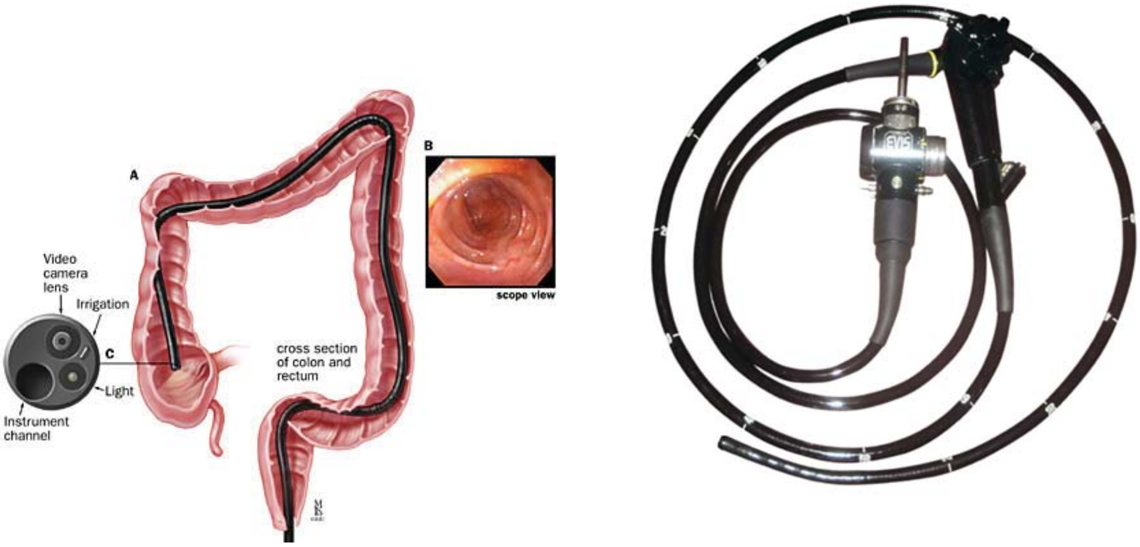
(g) Ulcerative colitis (h) Erosion (i) Melanosis

Figure 1.2: An inconclusive list of diseases that can be observed and diagnosed in GI tract [95]. These are the real images recorded from endoscopic equipment during routine examinations. Green box shows the status a colonoscope device.

scans and colonoscopy. However, in randomized trials, only endoscopic methods have shown precision enough to reduce CRC incidence.

There are several ways of detecting pathology in the GI tract, but currently available methods have limitations regarding sensitivity, specificity, access to qualified medical staff and overall cost. Here, the manual endoscopy, where the doctor inserts an endoscope in the patient, either via the mouth or the anus, is the recommended standard for detection and examination. An alternative to the manual colonoscopy (figure 1.3) is to perform the examination using a wireless camera pill, which is a video capsular endoscope (VCE) that can be swallowed by the patient and is able to record a video of the whole GI system.

However, scheduled testing (screening) of a population for a whole country is challenging due to high costs, a limited willingness by the patients to undertake the unpleasant procedure, high time consumption for the medical experts and a shortage of qualified medical personnel. Moreover, colonoscopy (the endoscopic examination of the colon) is unpleasant [143] for



(a) Colonoscopy (hopkinsmedicine.org) (b) Coloscope (olympus.com)

Figure 1.3: Colonoscopy is the endoscopic examination (a) of the large bowel and the distal part of the small bowel with a special type endoscope called coloscope (b) [116].

the patients, each requires about two staff-hours of medical personnel and often lesions are missed because of tiredness of the medical doctor or because a specific part in the colon was not reachable due to narrow passages in the colon. Furthermore, there are high costs related to these procedures. In the US, for example, colonoscopy is the most expensive cancer screening process with an annual cost of \$10 billion dollars [137], i.e., an average of \$1,100 per examination [138] (up to \$6,000 in New York). In the United Kingdom, the costs are around \$2,700 per examination [123]. Moreover, on average, 20% of polyps, precursors of CRC, are missed or incompletely removed, i.e., the risk of getting CRC depends mainly on the endoscopist's ability to detect polyps [69], thus requiring expensive specialized training for them.

To scale such examinations up to a large population either nationally or internationally, there are huge challenges that must be addressed to reduce cost per examination and to improve procedures for the detection of pathology (diseases). It is our vision that computer-based automatic execution of these tasks might be an important part of the solution, increasing the overall quality of the examinations and ultimately improving the patient outcome. The proposed technical solution targets ground-breaking research and innovation for global major health issues like colorectal, gastric and stomach cancer worldwide. By developing and studying an automatic system for the traditional push endoscopy and the modern VCEs, the aim is to make these examinations more easily accessible for patients and participants in screening programs, i.e., making the public healthcare system more scalable and cost-effective. Even more, we target utilization of the large amounts of disease records already store in the hospital information systems. Unfortunately, is not used [116] efficiently enough and holds a lot of potential, for example, by using it for efficient and accurate automatic analysis or by researching and developing live computer-assisted diagnosis based on it.

To summarize, the existing shortage of qualified medical personnel in conjunction with the high endoscopic procedures cost request for the computerization and automation of the complex

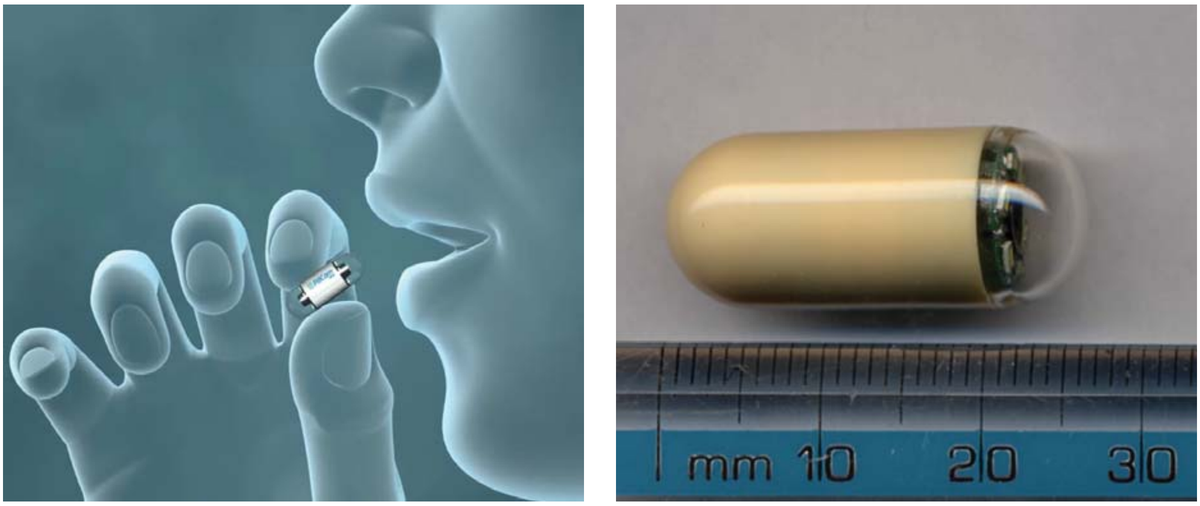

(a) Capsule endoscopy (igniteoutsourcing.com) (b) VCE (wikipedia.org)

Figure 1.4: Capsule endoscopy is a non-invasive procedure used to record internal images of the GI tract using a small swallowed VCE device equipped with a camera, a battery and a transmitting or recording module [116].

and labor-demanding GI tract diagnostic procedures allowing for assisted detection, highlighting and interpretation of lesions, diseases and findings in the GI tract in order to improve current medical practices and to save more lives.

1.2 Problem Statement

To satisfy the existing demands in assisted detection, highlighting and interpretation of lesions, diseases and findings in the GI tract via the computer-aided diagnostic procedures required to improve existing diagnostic practices and scale necessary GI tract examinations, we have started inter-disciplinary research of a next generation of the medical multimedia system, which will support endoscopists in the finding and interpretation of diseases in the entire GI tract.

The research question for this thesis is: ***Can modern computer vision and machine learning methods be used to build a holistic automated computer-aided diagnostic system supporting medical experts by analyzing images and videos in both live colonoscopy and VCE examinations?***

The goal of this thesis is to be a solid basement for building a complete, holistic and applicable medical multimedia system that can answer our research question and have a societal impact by helping people to survive lethal diseases. From our question, we define the objectives targeted by this thesis as follows:

Main Objective: Conduct research and develop a medical multimedia system that integrates and combines state-of-the-art tools with new and enhanced algorithms for detection and localization (highlighting) of pathological endoscopic findings and anatomical landmarks in the GI tract. The system should include the entire pipeline from content creation and annotation, learning and analysis to finally visualization of the output. The mechanisms should be combined in an extensible distributed architecture with real-time processing and efficient resource consumption for massive scale and high accuracy.

Sub-objective 1: Conduct research and develop a subsystem that can be used by the medical doctors (experts) to analyze, sort and annotate new and already collected images efficiently to minimize the amount of time required for such the annotations tasks. Additionally, search for the possibility to extract and make publicly available GI-tract-related medical imaging data already available in hospital medical information systems, with the following publishing datasets based on the annotated data.

Sub-objective 2: Conduct research and develop a subsystem for computer-based detection and decision support for live endoscopic procedures and VCE data analysis. The subsystem should receive video from endoscopic devices, perform analysis and show the clinicians both detected lesions and localization information overlaid over the main endoscopic video output. For the VCE case, the subsystem should be able to automatically analyze a large amount of VCE data in a reasonable time to enable future large-scale automatic population screening.

Sub-objective 3: Conduct research and develop a subsystem for visualization of the automatic detection results generated during live and VCE endoscopic examinations intended to decrease workload held by medical personnel during and after examination procedures.

To achieve these objectives, we teamed up with experienced specialists in the area of GI disease diagnosis to investigate how multimedia research can improve medical systems. In this thesis, we discuss and investigate why multimedia research is important and needed for the medical field and how a proper combination of medical experience, data collection, computer vision, deep- and machine-learning, automatic image and video analysis can become the key to solving medical challenges. Continuing from an initial version of the system called EIR developed earlier, this thesis presents the new, improved and extended version of the system called DeepEIR. The overall goal is to develop both, a live system assisting the visual detection and highlighting of different diseases during colonoscopies that are verified with different use cases, and a fully automated assisting system for the GI tract screening using VCEs, i.e., a small detached swallowable capsule-type device with one or more image sensors traveling along the GI tract. These aims come with strict requirements on the accuracy of the detection in order to avoid false negative findings (overlooking a disease). The live system should also avoid false positive findings (being too alarming can distract doctors and worry patients). Both systems should have low resource consumption and reasonable hardware requirements. The live-assisted system also must support real-time processing capabilities (defined [116] as being able to process at least 25 video frames per second (FPS)) captured with Full HD image quality, which is common for the modern endoscopic equipment. The screening-assisted system should be able to process a large amount of data and be able to adapt to a variety of used sensors characteristics from low-resolution to Full HD.

As the final outcome of this research, a holistic medical multimedia system is built for the GI endoscopy use case. Another outcome is an international cooperation of computer science researchers, medical experts and manufacturers of medical equipment already resulted in the problem-oriented work-groups, new datasets, medical protocols and disease atlases can also be used for the doctors' and IT researchers' training process. This cooperation is also going to continue the work after this PhD.

1.3 Scope and Limitations

Based on the research question and its objectives described in section 1.2, the scope of this thesis is on researching a complete medical multimedia system from annotation to visualization for the use case of different diseases and landmarks detection in the GI tract using mainly image and video data from different sources (traditional endoscopes and VCEs), and also prepare the algorithmic base of the system for other use-cases, including non-medical, and for the usage of various data types.

This research is the part of our larger project with the main goal of building a sale-ready medical information system that will support doctors in their daily duties. For this particular research, we limit the scope to the most common GI tract diseases, landmarks and findings, and two different medical data sources types. These scope limitations caused by the high complexity of the problem area and lacking of available data. High complexity is caused by the high variance of human diseases, their varying appearance, symptoms, localization and development stages, as well as limitations of diagnostic methods. The lack of available medical data is a well-known problem caused mostly by data privacy issues and the inability to use the data without explicit patient consent. This makes it hard to develop, evaluate and compare methods and algorithms. For testing, validation and evaluation, we used several publicly available datasets including our own newly collected datasets, which were made publicly available.

During this research, we faced with another limiting factor from the real world, which is the huge variety of the equipment used in different hospitals and even within single hospitals' departments. Different types of diagnostic equipment produce visual data with different resolution, color balance, sharpness, lighting conditions, frame rate, the field of view, quality, etc. The output of the equipment can be videos, still images, 360-degree images and videos, location information, etc. Even within a well-known group of our partner hospitals including ASU Mayo Clinic, Vestre Viken Hospital Trust, Rikshospitalet and the Karolinska University Hospital, the range of equipment includes multiple producers and different equipment models.

An additional limiting factor is the medical personnel's subjectivity and individual practice used in the data collection. There are no common standardized ways of collecting visual samples of diseases, and no well-documented strategies for the documentation of the diagnostic procedure, especially for GI tract medical interventions. This resulted in a wide variety of data collection practices and local standards used by different doctors. For example, in the Karolinska institute, doctors do not record videos at all and rely on extensive documentation using images. In Vestre Viken, medical experts store short video clips of the most important findings in combination with images. Even further, the availability of the already collected and annotated data in form of shared and publicly accessible datasets is very limited. This is addressed by introducing two newly collected, annotated and freely accessible public datasets created during this research in collaboration with the experienced doctors.

All these factors lead to strong requirements to the system adaptability and flexibility. The system developed with real-world cases in mind should be easily modifiable and able to adapt to different equipment used in different hospitals, different data formats and their properties, allow for handling of the individual data from each hospital if necessary.

Taking into account the limitations, the scope of this research should be reasonably limited. Our focus is on the detection of colon polyps, angioectasia flat lesion and bleedings. For these lesions, we provide frame-wise detection and point-wise localization (highlighting) via segmentation masks. We also provide detection for several normal findings and landmarks in the human GI tract. In order to be applied in real use-case scenarios, the system should be accurate, able to handle a large amount of data and be efficient in terms of processing speed.

1.4 Research Methods

In 1989, the ACM Education Board approved a report [45] created by a Task Force on the Core of Computer Science that determines and characterizes the structure of how research in computing should be approached. It defines computer science in its essence as an intersection between several central processes of applied mathematics, science and engineering. These central processes are basically reflected in the paradigms of theory, abstraction and design.

Theory is concerned with defining and characterizing the objects under study by formulating, hypothesize and determining possible relationships among objects, verifying relationship correctness and interpreting the results. *Abstraction* is used for modeling process and directly connected to experimental scientific methods. During the abstraction process, a researcher is investigating a problem, forming a hypothesis, creating a model, designing and running the experiments and, finally, collecting and analyzing the data. *Design* is tied with engineering and involves formulating of the requirements and creating appropriate solutions, followed by designing and implementing a system. This is concluded by the evaluation of the designed system.

For the theoretical part, the thesis touches elements of linear algebra, information theory, image and video representation, image processing with quality enchantment and color space operations, 2D vector-based geometric operations, building, training and testing of neural networks, human interpretation of multimedia content, etc. In the design of the algorithmic basis for the system, we developed a set of the complete end-to-end multi-purpose image classification and objects localization and segmentation algorithms.

To verify our hypothesizes, we created several experimental setups using different existing and newly collected datasets and did various experiments within our research group and public competitions in the relevant research communities. We explore image retrieval, analysis and features extraction techniques for single- and multi-class classification problems. We employ various image and multimedia data processing operations in different use cases. We study the performance of our system in terms of accuracy and processing speed aiming for real-world use cases and real-time applications. We also study the users' response to our solution and designed several user studies to collect annotation for the data and validate our system.

All the theories and abstractions presented in the thesis are implemented in several demo systems and prototypes. The developed software is thoroughly tested with the real data obtained from different equipment. The developed system was assessed by the experienced endoscopists from usability and efficiency points of view.

The developed system design is verified for technical correctness by creating various system prototypes for disease detection and localization that can be used in hospitals. To gain insights into domain-specific requirements, knowledge and to get access to actual medical data, we entry into a tight collaboration with experienced medical doctors from Vestre Viken Hospital Trust and Karolinska University Hospital.

The multi-purpose nature of the developed algorithms and complete parts of the system is verified by creating prototypes for objects detection on satellite images and out-of-patient medical images.

1.5 Contributions

The work presented in this thesis is a continued and extended research on the broad and complex topic of automated lesion detection in the human GI tract. The basic version of the EIR system was jointly developed by Michael Riegler and Konstantin Pogorelov, the author of this thesis. The basic EIR system was described in Riegler's thesis [112]. The second extended and improved version of the EIR system called DeepEIR is presented in this thesis. Both theses include the description of the background, motivation, problem, related work, algorithms and results obtained by Riegler and Pogorelov. The individual author's contributions are explained in chapter 5 and section 1.6.

The main contributions of this thesis are:

- technical development of a medical multimedia system called DeepEIR including annotation, detection, in-frame localization, visualization and proof-of-concept demonstration tools that confirm the potential of multimedia research in the health care system;

- broad comparison of various image classification approaches including classical machine learning and modern deep-learning-based approaches;

- research and development an efficient generalized distributed use-case-aware multimedia data processing method is able to achieve real-time performance for medical multimedia workload processing;

- demonstration and proof of the great potential of multimedia methods and experience of the multimedia community for applied research in medicine, and illustration how multimedia technology and methods can be used in the medical field to improve workflows, patient care and most importantly saving lives;

- contribution to the open-research community with the freely accessible novel open-source software libraries, datasets, prototypes and demos of the system;

- multiple published research papers about our findings and experiences.

Publications in top-tier conferences or journals support all the main contributions of the thesis. The diagram in figure 1.5 gives an overview of which of the attached papers contribute to which objectives. In more detail, the main contributions to the objectives defined in section 1.2 of the thesis are:

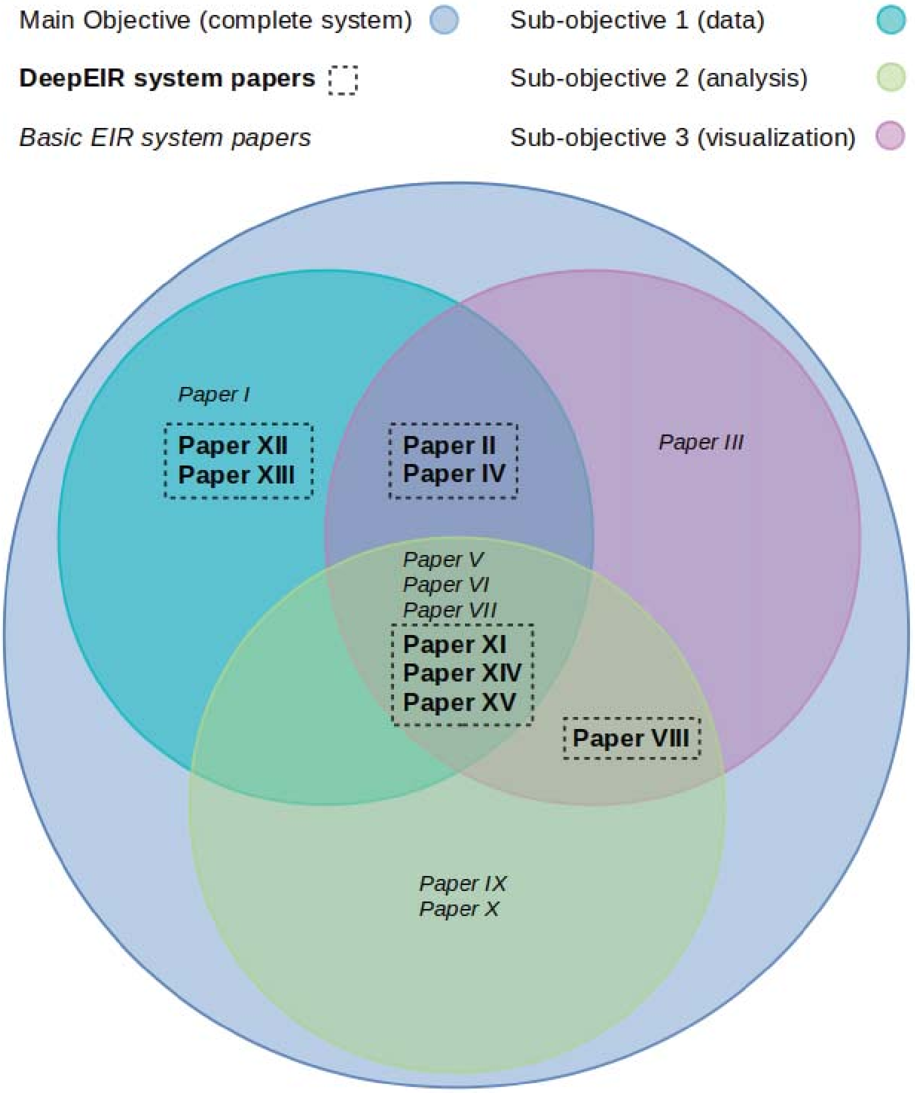


Figure 1.5: This diagram depicts the contributions for each of the in part II attached papers to the, for this thesis defined, objectives.

- **Contributions to the main objective:** We developed DeepEIR (the second version of the EIR system) for automatic detection and in-screen localization of lesions in the GI tract is capable for both real-time visual feedback during live colonoscopies using traditional endoscopic equipment and processing huge amount of data for population mass screening using VCEs [101, 102, 117, 118, 121].

  Using the ASU Mayo dataset [134], we showed that the detection subsystem of DeepEIR reaches high performance in terms of accuracy and processing. We can report a per-frame sensitivity and precision of almost 98% and 94%, respectively. This means that DeepEIR is able to find polyps in almost all cases with high precision. This can help the medical experts to save time and lives [101, 102, 117, 118, 121].

  Using the recent public Hospital Clinic of Barcelona dataset [23, 24] and our public datasets [94, 95], we showed that the detection subsystem of DeepEIR could reach high frame-wise classification performance in terms of accuracy, with a detection specificity of 94% and an accuracy of 90.9%. With the same datasets, the localization subsystem reaches the specificity and accuracy of 98.4% and 94.6%, respectively. The resulting performance of our detection and localization approaches is significantly higher than competing global-feature- and deep-learning-based approaches including the most recent real-time YOLOv2 [107] convolutional neural network (CNN).

Using the angiecstasia segmentation public dataset [23], we showed that the detection and the localization subsystems of DeepEIR can reach outstanding performance that exceeds clinical requirements (sensitivity and specificity higher than 85%). In summary, we achieved a sensitivity of 88% and a specificity of 99.9% for pixel-wise angiectasia localization, and a sensitivity of 98% and a specificity of 100% for frame-wise angiectasia detection [93].

Moreover, we compared DeepEIR with other existing systems and participated in a classification challenge where we showed that we outperform or reach at least same performance in accuracy as other state-of-the-art methods and that we are leading in terms of processing performance [25, 102, 117, 121]. Nevertheless, it is important to point out that the used datasets are still relatively limited in size and that evaluations on a large amount of data is recommended as soon as the data is available.

For the real-time processing challenge, we showed that DeepEIR can process at least 300 FPS for polyp detection, which is a good indicator that we created a scalable medical multimedia system able to process data in real-time [117]. We conducted research and implemented several ways of distributed and parallel processing by using heterogeneous computational architectures to improve the performance of the DeepEIR system. One of the methods that we investigated is the implementation of the detection and localization part on graphics processing units (GPUs) [101, 121]. Another method that we researched was to distribute the DeepEIR workloads via device lending [72, 102]. Both methods improved the processing performance significantly [72, 102].

We contributed to two open source projects: *Lire*, in the field of content-based image retrieval [80], and *OpenVQ*, on video quality [126]. We also released the base algorithm of DeepEIR as an open source project called Opensea [90].

For each part of the DeepEIR system, we developed working prototypes and demo applications. These prototypes and demo applications have been presented at conferences [17, 102, 117, 121]. All-in-all, we contributed with a holistic medical multimedia system for GI examinations [116] that will in the future help medical doctors to save lives.

- **Contributions to sub-objective 1:** For the annotation subsystem of DeepEIR, we conducted extensive research, together with our partner doctors, to make the process of medical knowledge transfer into our system easy and efficient for the medical experts. We explored and developed semi-supervised and cluster-based annotation tools [90, 98, 120].

  For medical data collection and publishing, we investigated the ethical and legal aspects of medical data use within our research process. We contacted several Norwegian hospitals and established relations with the data storage managing personnel. With the help of our medical-side collaborators, we made the agreements allowing us to extract and use the fully anonymized data from the hospital medical information systems. Using these data, we created two datasets (called Kvasir [95] and Nerthus [94]) and published them online freely accessible for educational and research purposes. We did our own evaluation of the datasets to give the baseline for other researchers [87, 99].

We used the published datasets for organizing Medico: The 2018 Multimedia for Medicine Task challenge within MediaEval Benchmarking Initiative for Multimedia Evaluation [61, 100, 119]. Our Medico challenge was accepted by the public and the research community. The datasets were evaluated by independent researchers and they are already used widely around the world.

- **Contributions to sub-objective 2:** As a basis for the detection subsystem, we developed a search-based classification algorithm that uses global image features, reaches good classification performance and is very fast at the same time [90]. As a basis for the localization subsystem, we developed a polyp localization algorithm based on the hand-crafted local features and global heat map post-processing, which reaches good polyp localization precision with reasonable high false-alert rate [25].

  We researched the problem of bleeding detection for VCE-captured videos and developed the basic bleeding detection and localization algorithm for the DeepEIR system [129].

  We implemented the multi-class global-features- and deep-learning-based classifiers are able to handle multiple lesions, landmarks and normal findings of the GI tract for the detection subsystem, investigated its efficiency both in terms of accuracy and processing speed and compared it to existing competitors [91, 96]. This formed a basis for developing the DeepEIR system into the holistic system that is usable and helpful in the real-world conditions.

  In order to extend the lesion detection capabilities of the DeepEIR system, we investigated and developed a GAN-based detection and localization approach for the angiectasia GI tract lesion [93]. Also, inspired by the success of our angiectasia detection approach, we researched and developed a GAN-based polyp detection and localization approach [92].

  We investigated the topic of deep neural network internal processes visualization for better medical image classification and classification understanding [62]. We investigated the tradeoffs using binary versus multi-class neural network classification for medical multi-disease detection [26].

  Based on the use cases addressed in the thesis and the DeepEIR system itself, we showed that the global- and local-feature-based algorithms together with the deep-learning-based approaches can form a strong basis for the multi-lesion detection system. We showed that the local hand-crafted features together with GAN-based approaches, can provide a good localization performance for the challenging lesions that are hard to see even for humans. In total, we proved that the developed algorithms are well suited to be applied in several use cases that involve image classification and analysis problems [91, 92, 93, 99, 101, 102, 116, 117, 118, 121].

- **Contributions to sub-objective 3:** We investigated different types of visualization for the output of the DeepEIR system. We developed the Web-based visualization application for research and medical experts [90] and its easier-to-use web-based version [121]. We developed an initial visualization approach that is able to visualize all outputs of the DeepEIR system [117], that was later developed in a live visualization application [96]. We investigated the problems of automatic reporting and developed a decision support system for deep-learning-based analysis in the medical domain [63, 64]

Additional contributions: Here, we list contributions that have been made during the PhD and are not related to the main topic of the thesis but were conducted because of it. These contributions are:

- We investigated and developed an approach to the flooding detection on the satellite images using our GAN-based approach that showed promising results [14, 15, 122] and built a unique system for collecting information and monitoring natural disasters by linking social media with satellite imagery can potentially save lives [13, 16].
- We investigated how context (a certain watching situation) influences the quality of experience for users when they are watching videos during a flight as a use-case. We hosted a MediaEval benchmark task [97] about this topic and published a dataset [115].
- We developed a system for efficient live and on-demand tiled HEVC 360 VR video streaming and investigated its performance in real use-case scenarios [55].
- We investigated and developed the new top-down saliency detection approach driven by visual classification, which showed promising performance on common saliency detection evaluation datasets [84].

1.6 This thesis author’s independent contributions

This thesis describes the DeepEIR medical multimedia system, which was built as the next step towards clinical-ready GI tract disease detection and localization computer-aided solution. This thesis author’s main independent contributions are the following:

- Speed optimization of the LIRE library used in the basic version of the detection subsystem (see Paper I).
- Development of the initial version of the global-feature-based clustering and visualization application (see Paper I).
- Development of the enhanced version of OpenSea classification tool used in the initial version of the detection system (see Paper II).
- Research and design of the efficient hyper-tree-based representation of the images clustering output (see Paper III).
- Development of hyper-tree-based visualization and annotation application has been used in data collection and annotation process (see Paper III).

- Research and design of the efficient feature extraction pipeline for the feature-based image classification approach used in visualization and detection subsystems (see Paper IV).

- Research and design of the real-time image-oriented database used in ClusterTag application (see Paper IV).

- Research and design of the real-time image clusters drawing module used in ClusterTag application (see Paper IV).

- Development of ClusterTag, the interactive visualization, clusterization and annotation application has been used in data collection and annotation process (see Paper IV).

- Research and design of the local hand-crafted-feature-based polyp localization approach. Development of the initial version of the localization subsystem using this approach (see Paper V).

- Research and design of the multi-CPU global features extraction. Development of the speed-improved feature-based version of the detection subsystem (see Paper V).

- Research and design of the GPU-accelerated features extraction. Development of the second version of the speed-improved feature-based detection subsystem (see Paper VI).

- Research and design of the GPU-accelerated speed-improved version of hand-crafted-feature-based polyp localization. Development of the second version of the localization subsystem (see Paper VI).

- Development of the detection and localization evaluation application for the MICCAI polyp finding challenge (see Paper VI).

- Research and design of the real-time detection and localization approach based on global and hand-crafted features. Development of the corresponding system evaluation application (see Paper VII).

- Research and design of the multi-class classifier for the detection subsystem. Development of the global-features- and deep-feature-based classification module for the DeepEIR system (see Paper VIII).

- Processing and annotation of the Kvasir dataset (see Paper VIII).

- Research and design of the second improved version of CUDA-based GPU-accelerated feature extraction and classification approach. Development of the corresponding module for the DeepEIR detection subsystem (see Paper IX).

- Research and design of the distributed multi-GPU feature extraction approach with the use of device landing for data processing speed improvement. Development of the corresponding parallel processing module and related DeepEIR detection subsystem modifications (see Paper X).

- Research of the pros and cons of the developed global- and deep-feature-based detection approaches. Detection and localization subsystems optimization for processing speed. Development of the live polyp detection and localization software (see Paper XI).
- Kvasir, Nerthus and Medico datasets preparation, annotation and publication. Development of the base-line classification algorithms for these datasets (see Papers XII and XIII).
- Research and design of the GI tract lesion segmentation approach (see Papers XIV and XV) based on a generative adversarial network (GAN) architecture.
- Research and design of the GAN-based pixel-wise localization and frame-wise detection approach for angiectasia and polyp lesions. Development of the new angiectasia and polyp modules for the detection and localization subsystems (see Papers XIV and XV).
- Research and design of the block-wise localization-via-detection approach for polyp lesions. Development of the additional polyp module for the detection and localization subsystems (see Paper XV).
- Research and design of the bladder cancer cells detection and localization approach (see subsection 3.6.4.1).
- Research and design of the spermatozoon detection and localization approach (see subsection 3.6.4.2).
- Performance evaluation of the EIR and DeepEIR systems in whole and their subsystems (see Papers I- XV).

In addition to the above contributions, the author also supervised several master students, organized workshops and was part of program committees for conferences. One of the latest papers describing author's GAN-based detection and localization approach (that was developed for the DeepEIR system) called "Deep Learning and Hand-crafted Feature Based Approaches for Polyp Detection in Medical Videos" won a Best Paper Award at the 2018 IEEE 31st International Symposium on Computer-Based Medical Systems [92] (Paper XV).

1.7 Outline

The research presented in this PhD thesis has been started from a simple medical image knowledge extraction task, which was rapidly developed into the whole and a complete end-to-end system is able to perform efficiently and to assist doctors during their routine work. From the very beginning, we decided to develop our system as a set of semi-independent subsystems, namely: annotation and data acquiring, analysis and visualization. We developed the corresponding methods and algorithms for these subsystems, finely tuned them for our use case and joined them into the complete DeepEIR system. Using our own and other publicly available data, we trained and evaluated our system, achieving promising results in terms of detection and localization accuracy. Finally, we investigated the system performance and successfully improved it reaching the goals of real-time (and even fasted) data processing performance and handling huge amount of data using distributed, parallel and GPU-enabled processing.

The rest of this thesis is organized as follows, giving an introduction to the main ideas that are described in more depth in the attached papers in chapter 5:

Chapter 2: Medical Multimedia Systems: We provide the background information about the human GI tract use case. We briefly describe the medical data challenges and our practical experience. We present related work focused on other medical multimedia systems, methods and datasets available.

Chapter 3: The DeepEIR System: We describe the complete DeepEIR system, its general overview, internal structure and connections to the outer world. Next, we describe the annotation, detection, localization and visualization subsystems and their algorithmic base, including some experimental results and discussion of real-world scenarios for the system. Then, we describe our experience with the system's data processing speed improvement, our approach to the real-time processing and handling of huge amounts of data. Finally, we describe our demos and prototypes that were used for testing and proving that the DeepEIR system can be used for the real-world medical use-case scenarios.

Chapter 4: Conclusion: We summarize and conclude this thesis and present ideas and concepts for further studies in the intersection between GI endoscopy and medical multimedia systems.

Chapter 5: Papers and Author's Contributions: Finally, we present all the core research papers that are included and discussed in this thesis. For each paper, we include a description of the author's contributions to it and indicate to which objectives it contributed.

Chapter 2

Medical Multimedia Systems

Medical multimedia systems introduce various challenges both for from research and development point of view. In this chapter we first look into the medical side of the problem area including the modern endoscopic devices. Then, we tackle the problem of medical data availability and search for available data sources. Next, we describe the state-of-the-art in medical data analysis and summarize currently unsolved challenges. Finally, we briefly describe our initial EIR system implementation and summarize our goals.

2.1 Gastrointestinal Tract Case Study

At a first glance, the modern health-care system is equipped with a huge amount of high-tech equipment to make the diagnosis, cure and follow-up processes fast, easy and convenient for the patients. In some areas, for example, blood sampling and computer tomography of internal organs, this is true. However, many of the medical investigations do still not only require a vast amount of preparation and manual work done by an experienced and specially trained doctor but also bring discomfort and pain into the patient's life.

Despite the progress in non-invasive human body scanning methods like, e.g., CT, MRT and ultrasound imaging, there are only few methods readily available for gastroenterologists for robust and reliable imaging of the GI tract and, especially, the upper part of digestive system and its colorectal area.

Upper Endoscopy An upper endoscopy is a procedure used to visually examine the upper digestive system with a tiny camera on the end of a long, flexible tube. A specialist in diseases of the digestive system (gastroenterologist) uses endoscopy to diagnose and, sometimes, treat conditions that affect the esophagus, stomach and beginning of the small intestine (duodenum). The medical term for an upper endoscopy is esophagogastroduodenoscopy. It can be done at a general practitioner's office, an outpatient surgery center or a hospital.

Colonoscopy A colonoscopy is an examination method used to detect changes or abnormalities in the large intestine (colon) and rectum. During a colonoscopy, a long, flexible tube (colonoscope) is inserted into the rectum. A tiny video camera at the tip of the tube allows the doctor to view the inside of the entire colon. If necessary, polyps or other types of abnormal tissue can be removed through the endoscope during a colonoscopy. Tissue samples (biopsies) can be taken during a colonoscopy as well. The most common reasons for colonoscopies include investigation of the GI-tract for signs and symptoms and possible causes of abdominal pain, rectal bleeding, chronic constipation, chronic diarrhea and other intestinal problems. Another reason is screening for colon cancer in people aged over 50, to be performed every ten years to screen for colon cancer. The previous history of colon polyps and CRC can also be a cause for necessary follow-up colonoscopies to look for and remove any additional polyps. This is done to reduce the risk of developing CRC.

Colonoscopy include conventional white-light endoscopy and virtual endoscopy [52]. Conventional white-light colonoscopy is regarded as the gold standard screening test for CRC [104]. Various randomized clinical examination and prospective cohort investigations have testified that conventional colonoscopy with polypectomy lowers the incidence of CRC significantly by 40-90% and decreases mortality [108]. Therefore, the demand for colonoscopy continues to increase.

Regardless of the achievement of colonoscopy in lowering cancer deaths, an important average miss rate for detection of both massive polyps and cancers is present and is approximated to be around $4 - 12$ percent [78, 88, 109]. The traditional endoscopies, such as colonoscopy and gastroscopy, only allow a physician to examine few regions of the GI tract. The traditional endoscopies cannot visualize the small intestine, due to cable length limitation. Furthermore, they can also tear intestinal walls in case of severe medical conditions, and endoscopies such as enteroscopy and push enteroscopy are uncomfortable for the patients. They are performed in real-time and are challenging to scale to a larger population [91]. Also, the procedure is expensive. In the United States, for instance, the colonoscopy is the most expensive cancer screening procedure with yearly expenses of 10 billion dollars, with an average of \$1,100 per person. In the UK, the prices are around \$2,700 per person. Norway has an average cost of about \$450 per examination. Scaling this to a population-sized cohort is very resource demanding and incurs enormous costs. Additionally, approximately one medical-doctor-hour and two nurse-hours, per evaluation is required that makes the real population-wide screening unrealistic scenario.

Prior to the introduction of wireless VCE, physicians could not examine the small intestine without any surgical operation. VCE was devised by a group of researchers in Baltimore in 1989, and afterwards introduced by Given Imaging Ltd., Yoqneam, Israel, as a commercial instrument. The device became publicly available in 2000 and used wireless electronic technology [67] that captures images of complete GI tract. This capsule-shaped pill can be swallowed by the patients in the presence of clinical experts without any discomfort [129]. Unlike conventional endoscopy procedures, this procedure investigates the entire GIT without pain, sedation and air insufflation. VCE has assisted more than 1.6 million patients worldwide until now. An additional advantage of this new technology is that the process of the physical examination that does not require sedation and is non-invasive, so it only applies little pain to the patient [41]. This entire VCE procedure enables clinicians to diagnose and detect ulcers, tumors, bleedings and other lesions in the small intestine to make offline diagnostic decisions afterward.

Moreover, GI tract inspection and screening is one of the areas under-covered by automation and computer-based support systems. Thus, the importance of corresponding GI-tract-oriented automated medical systems that provide support for diagnostics, examination, surgery, reporting and teaching cannot be underestimated. Moreover, regardless of the automation level, the support systems must be interactive, since the medical professionals must be in the loop to provide input, interpret and act on the results. Our investigation in the field showed that there is no complete medical multimedia system for analyzing multimedia data containing information about the GI tract in real-time. Thus, our primary goal is to develop such a complete system.

Following the general preconditions for medical system and common GI-tract-procedures, we define the following requirements:

1. Support for decision making during the traditional push enteroscopy (colonoscopy) and modern VCE.

2. Ability to process video streams from standard endoscopic equipment as well as images and videos captured by VCE.

3. Real-time detection and in-frame localization of different GI-tract diseases.

4. Ability to implement a complete processing pipeline including data collection, annotation, medical knowledge transfer, automatic analysis and visualization.

5. Ability to be extend to new diseases.

Up to now, detection of diseases in the GI tract was mostly focused on polyps. The main reason for this is the importance of polyp detection and the lack of well-annotated training and validation data for other gastric diseases. Automatic analysis of polyps in colonoscopies has been in the focus of research for a long time and several studies have been published. However, there are no complete systems, and none of the developed approaches can perform detection in real-time and support doctors by computer-aided diagnosis during colonoscopies. Furthermore, all the existing systems are limited to a very specific use case, trained and validated with very limited datasets or rely on a specific type of equipment.

2.1.1 Endoscopic Devices

As the first step in our research, we investigated the variety of the existing GI tract examination methodologies currently used in hospitals world-wide. All-in-all, we split them into two main categories: the indirect and direct investigation methods. Indirect methods include magnetic resonance imaging, various tomography, blood and fecal samples analysis. Direct methods are various endoscopic procedures and surgical interventions. In this research, we focus only on endoscopic diagnostic methods which give precise and reliable results with the reasonable cost and patient discomfort comparing to other methods. Also, comparing to, for example, fecal sample biomarker-based analysis, GI tract endoscopic screening covers all known GI-tract-related lesions. All types of endoscopic examination are performed using traditional and wireless video capsular endoscopic devices.

2.1.1.1 Traditional Endoscopes

Traditional endoscopy is a nonsurgical procedure used to examine a person's GI tract. It is performed using an endoscope, a flexible tube with a light and camera attached to it. The video stream is transmitted to an external TV monitor (and optionally a recording device and/or computer) showing the internal contents of a patient's GI tract. In general, the endoscopic procedures are non-invasive, but they can introduce a significant discomfort to the patient not only during the procedure itself, but also during a preparation phase. Most types of endoscopy require to stop eating solid foods for up to 12 hours before the procedure. Typically, the preparation requires strong laxatives or enemas to use the night before the procedure to clear the digestive system. There are many types of endoscopic procedures, but the most common are upper endoscopy and colonoscopy.

2.1.1.2 Wireless Video Capsular Endoscopes

Video Capsule Endoscopy (VCE) provides visualization of the gastrointestinal (GI) tract by capturing images or recording video using a small swallowed pill-like disposable capsule equipped with one or more cameras, a small processing device, memory or wireless transmitter and a battery. There are two main types of VCE capsules: Transmitting VCE (T-VCE) and Recording VCE (R-VCE).

The T-VCE capsule, also sometimes called Wireless Capsular Endoscope and Wireless Video Capsular Endoscope, performs capturing of images and immediately transmits video wirelessly from a capsule to a data recorder device worn by the patient. The T-VCE capsule is fully disposable and follows the swallow-and-forget concept that is convenient for both patient and doctor. The data captured becomes available for analysis and downloading almost instantly after activating and swallowing of the T-VCE capsule.

The R-VCE capsule performs capturing of images and stores the data on an onboard flash memory chip that eliminates the needs for a piece of additional external equipment on the patient's body. Instead, the R-VCE capsule requires recovering of the capsule from the patient's stool.

Both technologies have different pros and cons that make them suitable for different diagnostic and screening scenarios depending on the requirements in each specific case. Here, we describe them in short to demonstrate the potential of these technologies for the future discomfort-less examinations and national-wide screening programs.

The T-VCE equipment is often called Wireless Capsule System (WCS). It consists of 3 main components:

- a swallowed transmitting capsule endoscope device;
- a receiving and sensing system for receiving a data stream from the capsule, sensing pads or a sensing belt attached to the patient body, a data recording storage, and a battery pack;
- a workstation or personal computer with proprietary software installed and the interfaces to on-body module hardware.

All T-VCE capsule endoscope devices have similar components: a disposable plastic capsule, a complementary metal oxide semiconductor (CMOS) or high-resolution charge-coupled device (CCD) image capture system, a compact lens, a signal processing device, a wireless transmitter, white-light-emitting diode illumination sources, and an internal battery. Some modern capsules use magnetic and acceleration sensors to provide advanced localization information. The latest controllable capsules contain a magnet used to steer the capsule from outside of the patient's body.

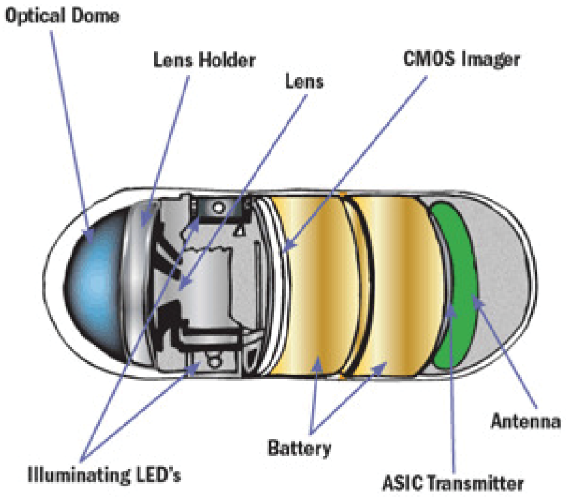


Figure 2.1: The internal components of wireless video capsule endoscope

The mode of data transmission is either via ultra-high frequency band radio telemetry or human body communications. The latter technology [77] uses the capsule itself to generate an electrical field that uses human tissue as the conductor for data transmission.

The first capsule model for the small intestine was approved by the US's Food and Drug Administration (FDA) in 2001. Over subsequent years, this technology has been refined to provide superior resolution, increased battery life, and capabilities to view different parts of the GI tract. Different producers provide a number of different T-VCE devices designed to be used for different parts of the GI tract, namely: small-bowel only, esophageal imaging and colon imaging.

Figure 2.1 shows the sample of the inner element of the T-VCE. This particular device is pill shaped ($26mm \times 11$mm), consists of light sources, a short focal length CCD camera, a transmitter of radio frequency and a few other electronic components. Once the capsule is swallowed by a patient, the WCE begins capturing images with 2-4 frame per second (fps) and sends them wirelessly to the recorder unit. This process produces between 50,000 and 80,000 images for each patient before the pill's battery is exhausted.

The R-VCE equipment is often called Storable Capsule Endoscope System (SCES). It consists of 3 main components:

- a swallowed recording capsule endoscope device;
- a data extraction system for obtaining recorded data from the capsule;
- a workstation or personal computer with proprietary software installed and the interfaces to the data extraction module hardware.

All R-VCE capsule endoscope devices have similar components: a disposable plastic capsule, a CMOS or CCD image capture system, a compact lens, a signal processing device, a large capacity onboard storage medium (several GB and more), white-light-emitting diode illumination sources, and an internal battery.

The recorded and compressed data is stored on the integrated storage medium, which can be done with a lower power consumption per recorded frame compared to wireless data transmission. This enables higher frame rates, better image resolution and longer recording time within capsules of the same size.

The main advantage of R-VCE is that patients do not need to wear an image recorder after swallowing the capsule, and they only need to be aware of the time of expelling the capsule from the body and collecting it. Currently [77], all R-VCE capsules require the excavation from the fecal masses, cleaning and the use of specialized communication module to extract the recorded data from the capsule. Nevertheless, compared to T-VCEs, screening using R-VCE devices can be performed virtually at any place (at home, at remote sites and on moving facilities, like ships and oil platforms), because only the capsule and simple and cheap disposable support equipment is required for the procedure.

After the data extraction, proprietary software is used to process and display the images in single or multiple views at any desired rates for R-VCEs and at rates of 5 to 40 frames per second for T-VCEs. Representative images and video clips can be annotated and saved. Most versions of available software have the ability to identify red pixels to facilitate detection of bleeding lesions. Localization of the capsule and monitoring of its movement through GI-tract are implemented for T-VCEs, but not yet for R-VCEs. Additional features include quick reference image atlases, and report generation capabilities. Different producers provide a number of different R-VCE devices designed to be used for different parts of the GI tract, namely: esophageal imaging, stomach imaging, small-bowel and colon imaging.

2.1.2 Medical Data

All described endoscopic devices generate a lot of multimedia data including still images, video streams, sensors and positioning data, etc. Some of this data is used only to provide real-time visual feedback to a doctor, some can be recorded locally or in hospital information systems for future use and reporting purposes. The access to such recorded data is strictly regulated by ethic and privacy grounds. From our experience, one of the most important challenges we meet during the development of the medical multimedia system is medical data availability and usability. Hospitals record, store and process a significant amount of data during routine procedures and patients' checks. This data contains information that is necessary for both efficient patient care and case investigation, and for educational and training purposes. However, the collected data is not used efficiently. This data holds much potential, for example, by using it for efficient and accurate automatic analysis or by researching and developing live computer-assisted diagnosis based on these generated data. Medical datasets also have the challenge that they usually contain many true negative examples, but not so many true positives. Furthermore, generalization is a vital ability for computer-assisted diagnostic systems that must be able to process data from different type of equipment (endoscope) used. Thus, a very important open question is how generalizable the proposed methods are.

During our research, we discovered only a few publicly and restrictively available datasets, which form a small set of reference images and video data can be used for the direct performance comparison of different approaches. Table 2.1 depicts the details of these datasets. As one can see, the available amount of data is relatively small, especially for the proper evaluation

Dataset Name	Data source	Frames contains example of	Dataset Size	Status	Description
CVC-ClinicDB [2]	Colonoscopy	Polyps	612 still images from 29 different sequences with polyp mask	Available	From 29 different sequences with polyp mask (ground truth)
ASU-Mayo Clinic Colonoscopy Video (c) Database [1]	Colonoscopy	Polyps	Training: 20 different videos Testing: 18 videos	Copyrighted	10 videos with polyps detections, 10 videos without polyps, GT available
CVC colon DB [3]	Colonoscopy	Polyps	300 frames with ROI	By explicit permission	15 short colonoscopy sequences (different studies)
ETIS-Larib Polyp DB [4]	Colonoscopy	Polyps	196 images	By request	196 images with GT
GI Lesions in Regular Colonoscopy Data Set [6]	Colonoscopy	GI lesions	76 instances	Available	15 serrated adenomas, 21 hyperplastic lesions, 40 adenomas
GastroAtlas [5]	Endoscopy	GI lesions	5,029 video clips	Available	Low-quality videos
The Atlas of Gastrointestinal Endoscopy [9]	Endoscopy	GI lesions	1295 images	Available	Esophagus, Stomach, Duodenum and Ampulla, Capsule Endoscopy, Inflammatory Bowel Disease, Colon and Ileum and some Miscellaneous
WEO Clinical Endoscopy Atlas [10]	Endoscopy	GI lesions	152 images	By explicit permission	One image per lesion
GASTROLAB [7]	Endoscopy	GI lesions	Several hundreds of images and several tenths of videos	Discontinued	Partially damaged/unavailable dataset
KID [8]	VCE	GI lesions	2,448 images and three videos	Discontinued, by request	Dataset access issues
Kvasir [95]	Various	GI lesions & landmarks	8,000 images, 8 classes, 1,000 images per class	Available, public, free for research and educational purposes	Our dataset. See section 3.1 for the description.
Nerthus [94]	Colonoscopy	GI findings	5,525 frames extracted from the 21 videos, 4 classes, from 500 to 2,700 frames per class	Available, public, free for research and educational purposes	Our dataset. See section 3.1 for the description.
Medico [100]	Various	GI lesions, landmarks and findings	14,033 images, 16 classes, from 4 to 2,331 images per class	Available, public, free for research and educational purposes	Our dataset. See section 3.1 for the description.

Table 2.1: Existing endoscopic image and video datasets

of the newly developed methods designed for real clinical setups. Also ground truth (GT) data for the available datasets is often missing or not accurate enough. Thus, in this research, we address this issue by introducing several new open-sourced and publicly available datasets.

2.2 Medical Image Analysis

The next naturally following question is how to use the endoscopic data efficiently both during live examinations to assist doctors, and later for automated diagnosis system development and medical personnel training. Widely used computer vision-based automatic visual data processing methods are designed for different use-cases and data types. Medical multimedia data analysis introduces a broad range of challenges mostly caused by the nature of the GI tract and nuances of the lesions that need to be detected, localized and assessed.

2.2.1 Challenges of Automatic Diseases Detection

Traditional colonoscopy and modern VCE offer an internal view of the digestive tract via non-surgical endoscopy technology. Following the progress in object recognition in the last few decades, computer-aided lesion detection methods have been in development with the ultimate goal of assisting doctors during routine procedures and lowering the lesion miss rate. However, automated lesion detection in live and recorded endoscopic video data is quite challenging because of the variation of polyps and other lesions inside the GI tract. GI tract findings can have color, texture and shape properties similar for different diseases and different for similar diseases in various stages of development. Findings can be covered by biological substances, such as seeds or stool, and lighted by direct and reflected light. Moreover, image coming from the endoscopic equipment can be interleaved, noisy, blurry because of lens defocus and camera motion, over- or under-exposed, it can contain static artifacts caused by lens contamination, borders, sub-images and a lot of specular reflections caused by the endoscope's light source. The GI tract can potentially have a wide range of lesions visible in endoscopy, as well as findings associated with benign/normal or man-made lesions. This phenomenon leads to a need for distinguishing between multiple classes of findings, including such with high level of visual similarity. In this scenario, both high precision and recall are of crucial importance, but also the frequently ignored system performance to provide live feedback because medical personal is assisted most efficiently while they perform the examination. Currently, there is no computer-assisted diagnosis or object recognition functionality implemented in endoscopic equipment for live examinations.

Modern VCE that has many advantages comparing to traditional push enteroscopy, require further improvement of the technology. Currently, clinicians must inspect 50,000 and more VCE images from between 4 and 12 hours of video footage to locate the diseases, which is a difficult task. They might miss the disease at an early stage due to visual fatigue or concentration loss. Moreover, VCEs do not have optimum lightning, making it more challenging to detect endoscopic findings in captured images than in images from traditional endoscopes. Also, during VCE procedures, the intestine is not inflated by injecting a small amount of low-pressurized gas into the GI tract via a endoscope, unlike in conventional endoscopy, where the expansion allows for precise and non-obfuscated images of the intestine walls. Nevertheless, ongoing research focuses at enhancing VCEs' hardware capabilities and at upgrading the techniques and algorithms developed for colonoscopies to work also for VCEs. While software developed by Given Imaging for VCE exists [74] and can detect active bleeding automatically, the sensitivity and specificity is very low, and no detection is implemented for other diseases at the moment. Moreover, the modern trends in multi-sensor VCE system design aims at the use-case where individuals can buy VCEs at the pharmacy and convey the video stream from the GI tract to the phone over a wireless connection. The video footage can be preprocessed on the mobile phone, in order to perform an initial analysis before the video footage is delivered to a processing back-end. In the best instance, the first screening results are accessible within eight hours after swallowing the VCE, which is the time taken by the camera to traverse the GI tract. Thus, the ability to execute and perform mass-screening of the GI tract relies on two fundamental research areas. First, it requires the improvement of a new generation of VCEs with better picture quality and the capacity to communicate with widely used mobile phones. Second, mass-screening demands a new generation of lesion detection algorithms able to process the captured GI tract multimedia data and video footage. Here, a preliminary analysis and task-oriented compressing of captured video footage before uploading into the cloud is of great significant because of the huge amount of data generated by VCEs.

2.2.2 State of the Art in GI Tract Lesion Detection

Early research on lesions detection in the human GI tract was mostly focused on polyp detection. The approach by Wang et al. [145, 146] was the most recent and best-working complete polyp detection system in the field of polyp detection when we started our system design and development. The system called Polyp-Alert employs edge-cross-section visual features and

Publication	Year	Detection Type	Sensitivity (Recall)	Specificity	Precision	Accuracy	F1	MCC	FPS	Dataset Size
Kang et al.[71]	2019	polyp / CNN	76.25	–	77.92	–	–	–	–	1187
Mori et al.[82]	2019	polyp / CNN	94	40	–	–	–	–	10	135
Byrne et al.[32]	2019	polyp / CNN	98	83	90	94	–	–	20	60,089
Urban et al.[140]	2018	polyp / CNN	96.8	95	–	96.4	–	–	98	8,641
Mori et al.[83]	2018	polyp / color, texture	92.7	89.8	93.7	–	–	–	2.5	61,925
Wang et al. [146]	2015	polyp / edge, texture	97.70	–	–	95.70	–	–	10	1.8m
Wang et al. [145]	2014	polyp / shape, color, texture	81.4	–	–	–	–	–	0.14	1,513
Mamonov et al. [81]	2014	polyp / shape	47	90	–	–	–	–	–	18,738
Zhou et al. [150]	2014	polyp / intensity	75	95.92	–	90.77	–	–	–	–
Li and Meng [75]	2012	tumor / textural pattern	88.6	96.2	–	92.4	–	–	–	–
Ameling et al. [20]	2009	polyp / texture	95	–	–	–	–	–	–	1,736
Cheng et al. [39]	2008	polyp / texture, color	86.2	–	–	–	–	–	0.076	74
Hwang et al. [66]	2007	polyp / shape	96	–	83	–	–	–	15	8,621
Alexandre et al. [18]	2007	polyp / color pattern	93.69	76.89	–	–	–	–	–	35
Kang et al. [70]	2003	polyp / shape, color	–	–	–	–	–	–	1	–
Riegler et al. [112]	2017	multi-class / global features	98.5	72.49	93.88	87.7	–	–	300	18,781

Table 2.2: A performance comparison of GI findings detection approaches. Not all performance measurements are available for all methods, but including all available information gives an idea about each method's performance. Also there are many done and ongoing research in the field, and this table present a selection of the most representative and recent results

a rule-based classifier to detect an edge along the contour of a polyp. The technique employs tracking of detected polyp edges to group a sequence of images in order to be able to detect the same polyp's appearances as one polyp event. The best achieved sensitivity of 97.70% and accuracy of 95.70% together with the relatively high processing speed measured as 10 FPS enabled initial clinical trials. We joined our research efforts recently resulting in the co-authored work [116]. However, the Polyp-Alert system is limited to the polyp use-case and it also does not provide low enough processing latency necessary for the live colonoscopies support.

Mamonov et al. [81] presented a simple polyp presence detection algorithm based on the geometrical shape of polyps and on the assumption that polyps often are hill-shaped objects bumped out of the surrounding tissue. With the main goal of reducing the number of frames that need to be manually inspected, the algorithm reached a sensitivity of 81.25% and a specificity of 90% for a per-polyp measure. For a per-frame measure only a sensitivity of 47% was reached with the specificity of 90.2%, which makes this detection algorithm not precise enough for real-time feedback generation.

Hwang et al. [66] developed a similar shape-based approach assuming that polyps are spherical or hemispherical geometric elevations on the surrounding mucosa. The method relies on a watershed-based image segmentation algorithm. Then ellipses are fitted into the segments by constructing a binary edge map for each segmented region using a least square fitting method. After the coarse size-based filtration, ellipses are further evaluated for matching of curve direction, curvature, edge distance and intensity. The interesting part of this approach is that after the first frame a potential polyp was detected, subsequent frames are also searched for the same characteristics using a mutual and information-based image registration technique. The method's evaluation showed reasonably high sensitivity and precision of 96% and 83%, respectively, achieving, at the same time, promising 15 FPS processing speed. Nevertheless, this and other shape-oriented approaches are strictly limited to polyp detection and cannot be easily extended to other flat or non-shaped diseases, e.g. bleeding, angioectasia, ulcers, etc.

The most recent works mostly incorporate modern CNN architectures as the detection and localization subsystems' basis. Mori et al. [82, 83] presented two complete polyp detection systems that were tested in real clinical trials. The first [83] system's detection algorithm is based on custom color and texture features extracted from every frame being processed with a following classification using a traditional ML-based SVM classifier. The system is able to process input frames at a rate of 2.5 FPS and has a corresponding sensitivity of 92.7%, specificity of 89.8% and precision of 93.7%. Despite the relatively high system performance, the overall data processing speed is not enough for convenient system use, due to an often limited polyp appearance time (a polyp can sometimes be clearly visible on one single frame in a 30 FPS video stream). The second [82] proposed detection system is based on a custom CNN architecture especially designed to work with a combination of traditional, magnified and narrow-band imaging (NBI) frames captured by a modern endoscopic system from Olympus. The developed system achieved a sensitivity of 94% and a specificity of 40% reaching near-real-time processing speed. Compared to many others, actual testing of this approach with real endoscopic equipment confirms the high quality of the designed software and the corresponding algorithmic base. Nevertheless, a test dataset with limited size was used for the system evaluation, rising the question of system's flexibility and ability to act in the different conditions. Moreover, the processing speed of 10 FPS is not enough for high-quality support during live colonoscopies. Moreover, both systems [83, 82] do not provide any localization information and are not able to highlight the polyp on a live view screen.

Byrne et al. [32] described an interesting Inception-based CNN architecture designed for NBI colonoscope imaging mode. With the ultimate goal of polyp detection, the detection algorithm provides a sub-class classification (hyperplastic polyp or conventional adenoma) of the found polyps. The performance numbers achieved on the validation set sized 18% of training set, are reported as a sensitivity of 98%, specificity of 83%, precision of 90%, and accuracy of 94%. The high measured method accuracy in conjunction with a relatively high processing speed of 20 FPS forms a solid basement for a complete detection system. However, the proposed detection method is suitable for NBI images only, which are normally used only after the actual polyp recognition by the performing endoscopist. Thus the method itself cannot be directly involved in a holistic polyp detection system.

Kang et al. [71] developed a novel approach based on two joint Mask R-CNNs based on the pre-trained ResNet50 and ResNet101 models. A bit-wise combination of the output masks used to enhance the segmentation performance of the proposed method is able to provide not only detection output, but also a precise polyp localization mask within an input image. With the pixel-wise sensitivity of 76.25% and precision of 77.92% this method demonstrates a promising potential for future complete lesion detection, but it requires significantly wider evaluation on the various datasets, as well as the corresponding processing speed testing.

Urban et al. [140] presented a set of custom CNN architectures especially designed for the dual binary detection and regression localization modes. The primary polyp recognition is implemented by a combined CNN model performing the optimization of the polyp size and location with mean-squared error loss; optimizing the overlap between the predicted bounding box and the ground truth; and a variation of the "you only look once" (YOLO) algorithm, in which the CNN produces and aggregates multiple individual weighted predictions of polyp size and location in a single forward pass. Authors tested randomly initialized and well-known ImageNet-pre-trained models. The best performing model incorporates initial weights from the ResNet50 network, and was able to reach an accuracy of 96.4%, sensitivity of 96.8% and specificity of 95%. The top processing speed was measured as 98 FPS on a high-end consumer-grade PC equipped with the recent GPU. However, the stated higher-than-real-time processing speed was reached for low-resolution 224x224 pixels input images and can potentially lead to a high miss rate for small polyps.

All-in-all, the state of the art methods and existing complete systems show the great potential of computer-based lesions detection in the human GI tract. Existing solutions can not only reach high performance in terms of accuracy, sensitivity, specificity and precision, but also demonstrate real-time or near-to-real-time data processing capabilities. However, despite the achievements of the different research teams in the last 5 years, there is still a lack of a complete holistic automated computer-assisted decision-making-support system that can perform well both during live endoscopic procedures and a posteriori VCE-captured imaging data analysis. Moreover, none of the existing complete systems can detect multiple diseases simultaneously and provide a live feedback to the endoscopists with both multi-class detection and detected lesion localization. With the work conducted in this thesis, we have beaten the mentioned problems and provided the medical society with a ready-to-use solution for GI-tract abnormality detection and localization.

2.2.3 Basic EIR System: The Proof-of-Concept

Our first EIR polyp-only detection system presented in Riegler et al. [112] is based on non-CNN image processing principles. The detection subsystem analyzes multimedia data, such as videos and images. All the frames processed by the detection subsystem are separated into two positive and negative classes. Two sets containing example images for abnormalities and images without any abnormality are used as the model for the disease detector. Global image features from Lire [79] library are used to compare images in the search-based two-class classification algorithm. The basic localization subsystem implements a model for polyp localization using a hand-crafted object localization method, based on the geometrical shape of polyps. We evaluated our first version of the EIR system using publicly available datasets. The experimental evaluation showed EIR's promising detection efficiency with the following performance metrics: a sensitivity of 98.5%, a specificity of 72.49%, a precision of 93.88% and a accuracy of 87.7%. Polyp localization performance evaluation showed a precision of 28.7% and a sensitivity of 76.1%.

2.3 Summary

It seems that despite of a rapid development of the new medical devices, complete medical multimedia systems are not in focus of active research, nor main-stream development. Most medicine-oriented research is now focused on algorithms, especially deep-learning-based, for the detection of diseases, not on complete medical systems design and implementation. Even further, the widely presented different lesion recognition approaches that are positioned as having a high performance properties are, in fact, very narrow and focused on one exact lesion or have been trained and evaluated using small private datasets preventing any reproducibility and cross-evaluation attempts. The only few examples that focus on more than one component seem to ignore data processing speed and real-time performance problems, or do not reach the use-case-dictated performance requirements. Most of the modern approaches incorporate various deep learning techniques, which is a hot and promising direction in the field of medical image processing, but requires a large amount of well-annotated training data that can be problematic in the highly-privacy-restricted medical scenarios.

To address these problems, in 2015, the development of a complete medical multimedia system with real-time and applied use-cases in-mind was started. The very first version of the EIR system incorporates our search-based classification approach that was presented by Michael Riegler in his PhD thesis [112], which demonstrated promising results and promised further potential for our use-case of disease detection in the GI tract. This thesis presents a further development of the complete EIR system, introducing new algorithmic and deep-learning-based detection, localization and segmentation approaches. Together with the newly collected and published open-source datasets, developed annotation, visualization and high-performance processing subsystems, the new DeepEIR system reaches the goal of a holistic medical decision-support system. To the best of our knowledge, the medical multimedia system developed and described in this thesis is the first system that reaches total flexibility and extendability in terms of diseases and objects that can be detected, localized and segmented, and, at the same time, provides the outstanding data processing performance with a proper and comparable evaluation of its performance with newly collected, annotated and published datasets.

In the next chapters, we present our holistic and complete medical multimedia system and all the sub-components. We also present our open-source datasets. Furthermore, we show a complete evaluation of the system performance in terms of accuracy with different GI tract findings and data processing speed including our heterogeneous and distributed improvements of EIR system.

Chapter 3

The DeepEIR System

Our primary practical objective is to develop a system that will support doctors in GI tract disease detection during both traditional live endoscopies and modern VCE procedures including home- and hospital-based wide population screening. Thus, the system must:

- be easy to use and less invasive for the patients than existing methods;
- support multiple classes of detected GI diseases, objects and landmarks;
- be easy to extend to new diseases and findings;
- handle multimedia content in real-time and process at least 30 FPS for Full HD videos;
- be designed and tested for live real-time computer-aided diagnosis;
- achieve high classification performance with minimal false-negative classification results;
- have a low computational resource consumption;
- be able to process huge amounts of pre-captured data;
- support scaling, parallel and distributed processing.

Implementation of these properties provide an efficient system allowing for a reduced number of specialists required for a larger population coverage with GI tract investigation, and dramatically increased number of users potentially willing to be screened.

The second extended and improved version of EIR system is called DeepEIR (see figure 3.1) and was designed with all mentioned properties in mind. It consists of three main parts: the data acquisition, preparation and annotation subsystem, the automatic analysis subsystem and the visualization and computer-aided diagnosis subsystem. The main DeepEIR's "brain" - the analysis subsystems is designed in a modular way to be easily extended to new diseases or subcategories of diseases, as well as for other not-implemented-yet tasks like size determination, 3D shape recognition, etc. Currently, we have implemented two types of analysis subsystems: the detection subsystem that detects different irregularities in video frames and images, and the localization subsystem that localizes the exact position of the disease within the frame. The detection subsystem is designed to only determine the presence of an irregularity within the frame.

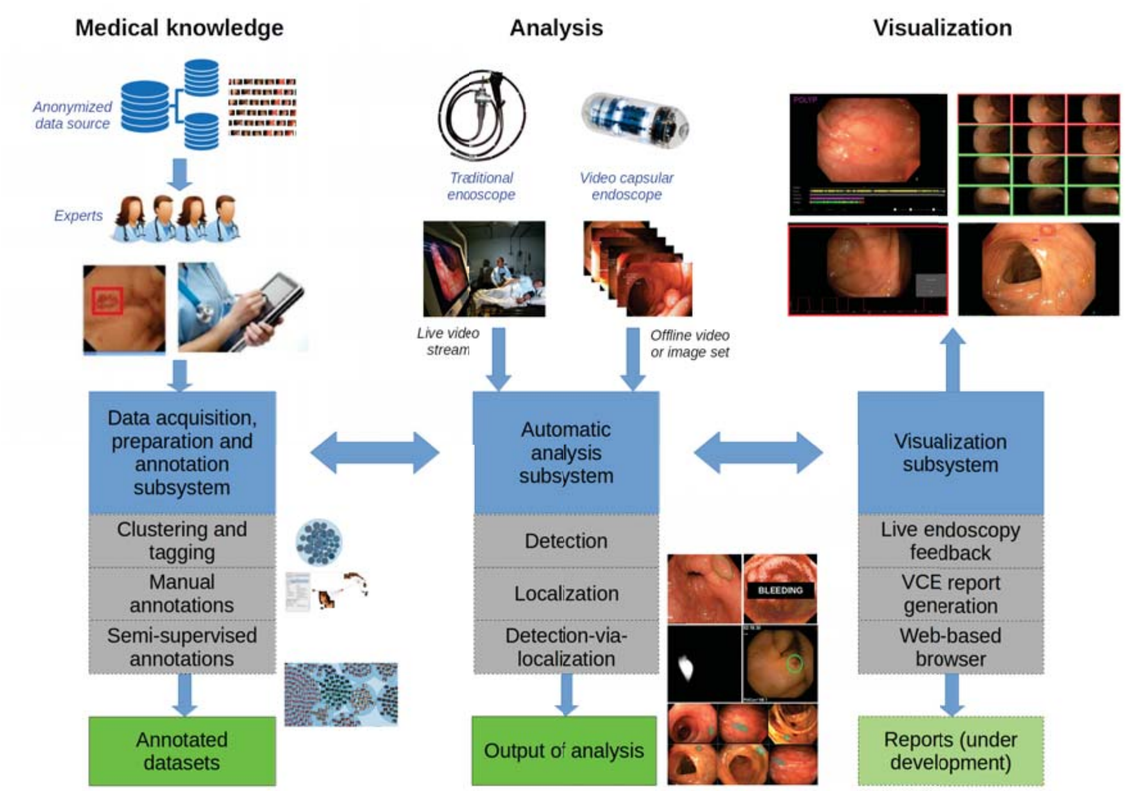


Figure 3.1: A complete overview of the DeepEIR system. The system consists of data acquisition, preparation and annotation, automatic analysis and visualization subsystems.

The exact position of the detected object is determined by the localization subsystem. Each detection subsystem therefore can be accompanied by the corresponding localization subsystem. The localization subsystem can be also implemented in two different ways. One uses the output of the detection system as input and processes only frames marked as containing a localizable disease. Another can act as the primary analysis agent and can perform frame segmentation with a following localization and detection-via-localization operations.

3.1 Data Collection

Despite automatic detection of diseases by use of computers is a life-saving area of applied science, it is still an under-explored field of research not only because of the absence of the well-performing algorithms and analytical models, but also because of a significant lack of data available for analysis, training and evaluation of the automatic methods being developed. Datasets containing medical images are hardly available, making reproducibility and comparison of approaches almost impossible. Thus, as a vital part of our research, we aimed also at the collection and annotation of an adequate and big enough dataset that can be used not only in this particular research, but that can also contribute to the research community and positively impact the current research comparability. We achieve this by collecting medical data, sorting and annotating it, publishing related papers with suggested common metrics, and the preliminary evaluation of results of the different classification methods, and, finally, making the datasets publicly available and free for non-commercial, educational and research purposes. Our public datasets contain images from inside the GI tract, and by providing them, we hope to invite and enable multimedia researchers into the medical domain of detection and retrieval. Moreover, by our public datasets, we especially address the common problem of research comparison when the results are hard to reproduce due to a lack of publicly available medical data.

3.1.1 Privacy, Legal and Ethics Issues

It is almost impossible to just obtain medical data from relevant medical institutions and hospitals for research purposes. All medical data is considered personal data and, therefore, is strongly protected from unauthorized use and distribution. That is most probably the main reason of lack of datasets that are publicly available, compared to traditional computer vision and information retrieval tasks. During this research, as a first and the most important challenge, we solved this problem by entering into a wide collaboration with a number of Norwegian hospitals and research teams working there. In order to get permission to download, process and publish the medical data, particularly image data from GI tract, we performed a detailed investigation into current Norwegian regulations concerning medical data privacy and possible ways to obtain a massive amount of data with respect to the data protection laws. As the result of these activities, we entered into an agreement with Vestre Viken Hospital Trust, allowing our research team to download anonymized data from hospital information systems and transfer it using secure media to our research facility. Than, we performed an additional data check and purification in order to fully remove any data can be used for potential patient tracking and deanonymization including removal of time stamps and EXIF information from the media files. As a negative consequence caused by the full data anonymization, we have lost all information that can help us to automatically classify the obtained raw data into relevant classes. Thus, next, we performed sorting and classification of the raw data. Due to a significant shortage of free time among the collaborating medical personnel, we decided to focus first on still images, leaving the captured video clips for the next project stages. The GI tract images were carefully annotated by one or more medical experts from Vestre Viken Hospital Trust and the Cancer Registry of Norway. In addition, a subset of the colorectal videos was annotated by a number of medical experts from Norway, Sweden, UK, US and Canada through a web based system. All the annotated images and videos will be released as an addition to the already published datasets regarding the specific use-cases assessments.

3.1.2 Sources of the Data

The raw data itself is collected using endoscopic equipment at Norwegian Vestre Viken Hospital Trust, which consists of 4 hospitals and provides health care to 470.000 people. One of these hospitals (the Bærum Hospital) has a large gastroenterology department from where training data have been collected and will be provided, making the dataset larger in the future. The Cancer Registry of Norway provides new knowledge about cancer through research. It is part of South-Eastern Norway Regional Health Authority and is organized as an independent institution under Oslo University Hospital Trust. The Cancer Registry of Norway is responsible for the national cancer screening programmes with the goal of preventing cancer death by discovering cancers or pre-cancerous lesions as early as possible.

3.1.3 Created Datasets and Reproducibility

We published three medical datasets called Kvasir, Nerthus and Medico, and a number of subversions. Kvasir and Nerthus are general-purpose datasets that can be directly used for building and evaluation of medical image recognition, information retrieval, single- and multi-class classification algorithms. Medico is a special-purpose dataset built based on Kvasir, and it is used in our Medico: The Multimedia for Medicine Task, which is part of a wider MediaEval Benchmarking Initiative for Multimedia Evaluation. All the datasets are publicly available online, and we evolve them constantly by adding new images and image classes.

3.1.3.1 Kvasir

The Kvasir dataset is our main contributing dataset representing a collection of images from different parts of the human GI tract. It consists of images, annotated and verified by medical doctors (experienced endoscopists), including several classes showing anatomical landmarks, pathological findings or endoscopic procedures in the GI tract. It contains hundreds of images for each class. The number of images is sufficient for different tasks, e.g., image retrieval, machine learning, deep learning and transfer learning, etc. The dataset is made up of the images of anatomical landmarks, pathological findings (lesions) and their removal procedures as well as a variety of normal GI findings. The anatomical landmarks include Z-line, pylorus and cecum, while the pathological findings include esophagitis, polyps, ulcerative colitis. In addition, we provide several set of images related to the removal of lesions, e.g., "dyed and lifted polyp", the "dyed resection margins", etc. The normal findings include various types of normal colon wall tissue and a variety of stool and food leftovers that can be observed during colonoscopies.

The dataset consists of images with resolution from $720x576$ to $1920x1072$ pixels and is organized in a way where images are sorted in separate folders named accord to their content. Some of the included classes of images have a green picture in picture illustrating the position and configuration of the endoscope inside the bowel, by use of an electromagnetic imaging system (ScopeGuide, Olympus Europe) that may support the interpretation of the image. This type of information may be important for later investigations and it is thus included, but it must be handled with care for the detection of the endoscopic findings.

Lesions

A pathological finding (lesion) in this context is an abnormal feature within the gastrointestinal tract. From the endoscopic point of view, it is visible as a damage or change in the normal mucosa. Finding may be a sign for an ongoing disease or a precursor to cancer. Detection and classification of pathology is important in order to initiate correct treatment and/or follow-up of the patient. The most common and dangerous findings include colon polyps, colorectal cancer, gastrointestinal bleedings, angioectasia, esophagitis, and ulcerative colitis.

Colon Polyps

Polyps are lesions within the bowel that are detectable as mucosal outgrows. An example of a typical polyp is shown in figure 3.2(a). The polyps are either flat, elevated or pedunculated, and can be distinguished from normal mucosa by color and surface pattern. Most bowel polyps are harmless, but some have the potential to grow into cancer. Detection and removal of polyps are therefore important to prevent the development of colorectal cancer. Since polyps may be overlooked by doctors, automatic detection would most likely improve examination quality. The green boxes within the image show an illustration of the endoscope configuration. In live endoscopy, this helps to determine the current localisation of the endoscope-tip (and thereby also the polyp site) within the length of the bowel. Automatic computer-aided detection of polyps would be valuable both for diagnosis, assessment and reporting.

Polyp Removal

Polyps in the large bowel may be precursors of cancer and are therefore removed during endoscopy if possible. One of the polyp removal techniques is called endoscopic mucosal resection. This includes injection of a liquid underneath the polyp, lifting the polyp from the underlying tissue. The polyp is then captured and removed by use of a snare. Lifting minimizes risk of mechanical or electrocautery damage to the deeper layers of the GI wall. Staining dye (i.e., diluted indigo carmine) is added to facilitate accurate identification of the polyp margins. Computer detection of dyed polyps and the site of resection would be important in order to generate computer aided reporting systems for the future.

Figure 3.2(b) shows an example of a polyp lifted by injection of saline and indigocarmine. The light blue polyp margins are clearly visible against the darker normal mucosa. Additional valuable information related to automatic reporting may involve successfulness of the lifting and eventual presence of nonlifted areas that might indicate malignancy.

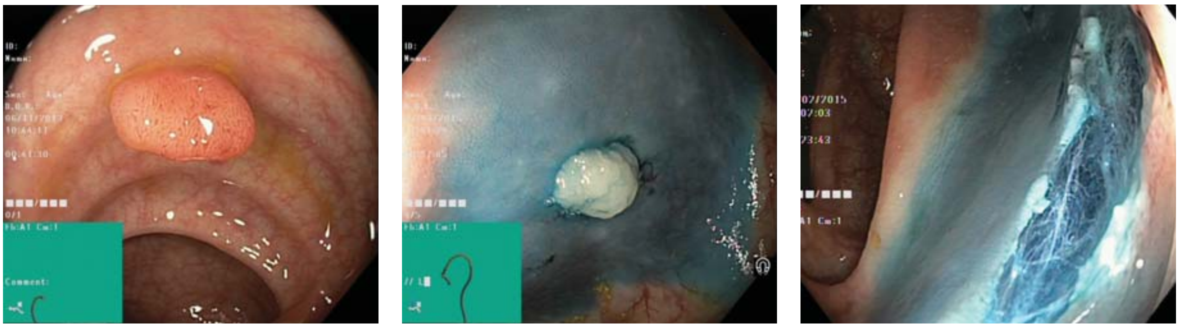
The after-removal resection margins are important in order to evaluate whether the polyp is completely removed or not. Residual polyp tissue may lead to continued growth and in the worst case malignancy development. Figure 3.2(c) illustrates the resection site after removal of a polyp. Automatic recognition of the site of polyp removals is of value for automatic reporting systems and for computer aided assessment on the completeness of the polyp removal.

Esophagitis

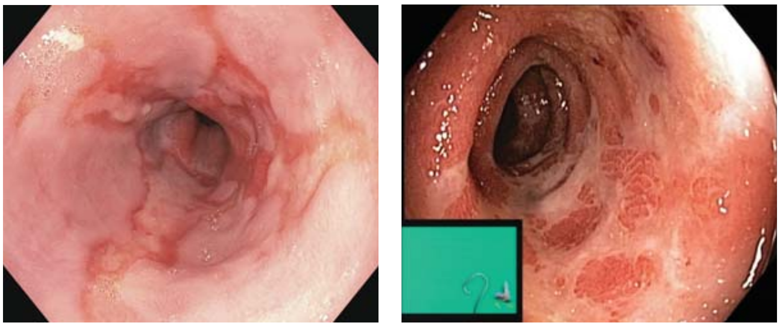
Esophagitis is an inflammation of the esophagus that is visible as a break in the esophageal mucosa in relation to the Z-line. Figure 3.2(d) shows an example with red mucosal tongues projecting up into the white esophageal lining. The grade of inflammation is defined by the length of the mucosal breaks and proportion of the circumference involved. This is most commonly caused by conditions where gastric acid flows back into the esophagus as gastroesophageal reflux, vomiting or hernia. Clinically, detection is necessary for initiating treatment to relieve symptoms and prevent further development of possible complications. Computer detection would be of special value in assessing the severity and for automatic reporting.

Ulcerative colitis

Ulcerative colitis is a chronic inflammatory disease affecting the large bowel. The disease may have a large impact on the quality of life, and diagnosis is mainly based on colonoscopic findings. The degree of inflammation varies from none, mild and moderate to severe, all with different endoscopic aspects. For example, in a mild disease, the mucosa appears swollen and red, while in moderate cases, ulcerations are prominent. Figure 3.2(e) shows an example of ulcerative colitis with bleeding, swelling and ulceration of the mucosa. The white coating visible in the illustration is fibrin covering the wounds. As mentioned earlier, an automatic computer aided assessment system will contribute to more accurate grading of the disease severity.



(a) Colon Polyp (b) Inked and lifted polyp (c) Polyp removal resection



(d) Esophagitis (e) Ulcerative colitis

Figure 3.2: Sample images of the GI tract lesions included in the Kvasir dataset.

Anatomical Landmarks

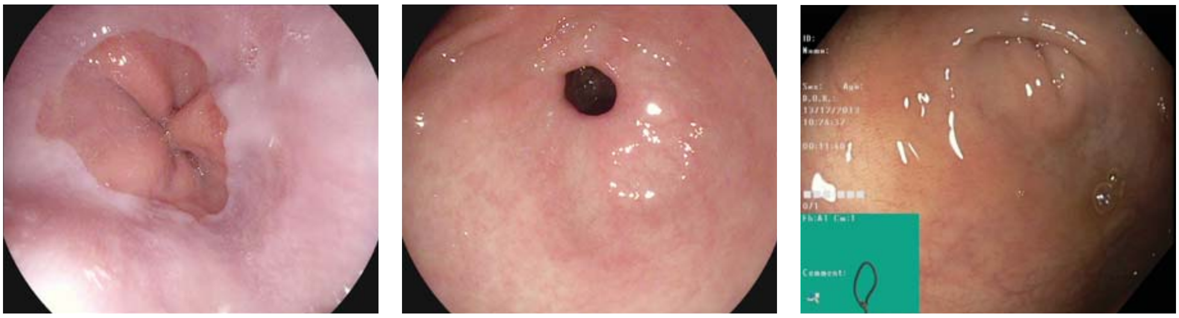
An anatomical landmark is a recognizable feature within the GI tract that is easily visible through the endoscope. Landmarks are essential for navigation and as a reference point for describing the location of a given finding. The landmarks are also be typical sites for pathology like ulcers or inflammation. A complete endoscopic rapport should preferably contain both brief descriptions and image documentation of the most important anatomical landmarks [111].

Z-line

The Z-line marks the transition site between the esophagus and the stomach. Endoscopically, it is visible as a clear border where the white mucosa in the esophagus meets the red gastric mucosa. An example of the Z-line is shown in figure 3.3(a). Recognition and assessment of the Z-line is important in order to determine whether a disease is present or not. For example, this is the area where signs of gastro-esophageal reflux may appear. The Z-line is also useful as a reference point when describing pathology in the esophagus.

Pylorus

The pylorus is defined as the area around the opening from the stomach into the first part of the small bowel (duodenum). The opening contains circumferential muscles that regulates the movement of food from the stomach. The identification of pylorus is necessary for endoscopic instrumentation to the duodenum, one of the challenging maneuvers within gastroscopy. A complete gastroscopy includes inspection on both sides of the pyloric opening to reveal findings like ulcerations, erosions or stenosis. Figure 3.3(b) shows an endoscopic image of a normal pylorus viewed from inside the stomach. Here, the smooth, round opening is visible as a dark circle surrounded by homogeneous pink stomach mucosa.



(a) Z-line (b) Pylorus (c) Cecum

Figure 3.3: Sample images of the GI tract landmarks included in the Kvasir dataset.

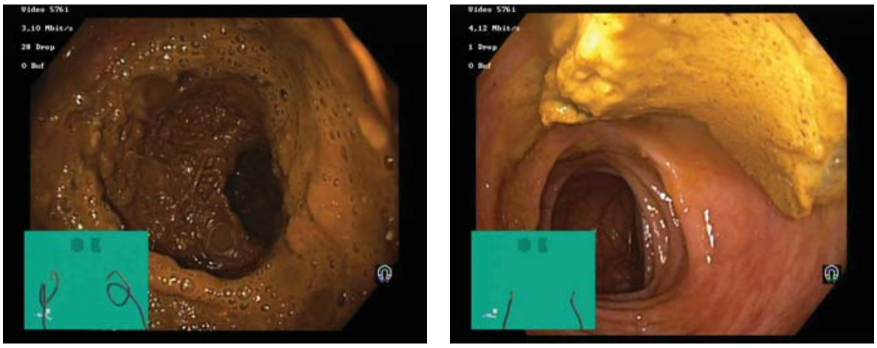
Cecum

The cecum is the most proximal part of the large bowel. Reaching the cecum is the proof for a complete colonoscopy [21]. Therefore, recognition and documentation of the cecum is important. One of the characteristic hallmarks of the cecum is the appendiceal orifice. This, combined with a typical configuration on the electromagnetic scope tracking system, may be used as proof for cecum intubation when named or photo-documented in the reports [110, 141]. Figure 3.3(c) shows an example of the appendiceal orifice visible as a crescent-shaped slit, and the green picture in picture shows the scope configuration for the cecal position.

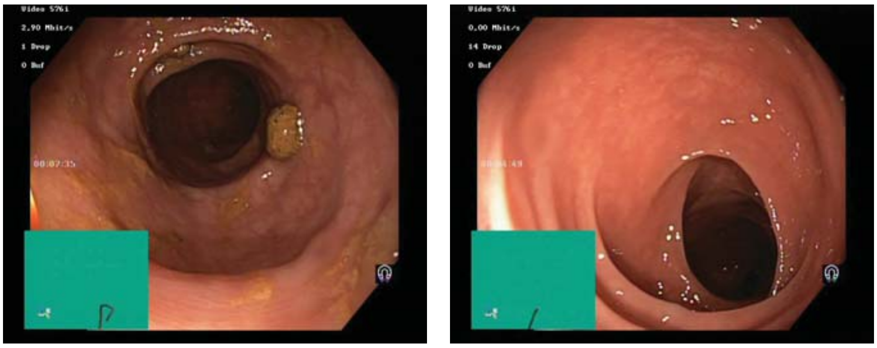
3.1.3.2 Nerthus

The Nerthus dataset is an auxiliary dataset addressing an important problem of adequate GI tract preparation which is a required pre-condition for the successful colon investigation and treatment. Traditionally, the bowel preparation quality has been categorized as poor, adequate or good. Such classification of bowel cleanliness often lacks clear definitions, and the judgement on quality tends to be subjective. This may result in significant inter-observer variation. To minimize the inter-endoscopist variation, new score-based methods of assessing bowel cleanliness have been introduced during the last decade. The state-of-the-art scoring system that is probably the best validated and most frequently used scoring system in both routine clinic and screening settings today is called the Boston bowel preparation scale (BBPS). It divides the bowel into three sections (right, middle and left) and scores the bowel cleansing within each section according to a defined numeric scale. It uses only a four-point scoring system (ranges from 0 to 3). Despite a promising standardization potential, there is no publicly available dataset can be used as a gold standard and a reference set for medical personnel training.

The Nerthus dataset consists of 21 videos with a resolution of $720x576$ with a total number of $5,525$ frames, annotated and verified by medical doctors (experienced endoscopists), covering 4 classes that show the four-score BBPS-defined bowel-preparation qualities. The number of videos per class varies from 1 to 10. The number of frames per class varies from 500 to $2,700$. The number of videos and frames is sufficient to be used for different tasks, e.g., image retrieval, machine learning, deep learning and transfer learning, etc. The videos are sorted into separate folders named according to their BBPS-bowel preparation quality score (see figure 3.4 for the examples). Most of the included videos and images have a green picture in each frame, illustrating the position and configuration of the endoscope inside the bowel. This is obtained



(a) BBPS 0 (from splenic flexure) (b) BBPS 1 (from descending colon)



(c) BBPS 2 (from sigmoid colon) (d) BBPS 3 (from rectum)

Figure 3.4: Sample images for each bowel preparation ("cleanliness") score according to BBPS.

from an electromagnetic imaging system (ScopeGuide, Olympus Europe) and may support the interpretation of the image. This type of information may be important for later investigations on segmental position within the bowel.

3.1.3.3 Medico Task

The Medico: Multimedia for Medicine Task is an image recognition and classification challenge running for the several years as a part of MediaEval Benchmarking Initiative for Multimedia Evaluation. It focuses on detecting abnormalities, diseases, anatomical landmarks and other findings in images captured by medical devices in the GI tract. The task provides to the participants a detailed use-case description, including its importance and related challenges, the dataset with the ground truth, the description of the required runs and the evaluation metrics. The task introduces a lot of challenges related to correct medical image classification as well as the related lesion localization and differentiation. The task has repeatedly used the latest version of the task's dataset, now consisting of more than $10,000$ images, which are annotated and verified by experienced endoscopists.

The whole dataset is split into two equally sized development and test datasets. Pre-extracted visual features for all the data are also provided. The ground truth for the data is collected from the medical experts annotations. Both the development and the test datasets consist of images sorted into classes with different numbers of images in each class stored in two archives: image archive and features archive.

The image archive contains raw images sorted into classes with different number of images per each class. In the development dataset, the images are stored in separate folders named according to the name of the class images that belongs to. In the test dataset, all the images are stored in one folder. The images of the dataset come from equipment installed in Norwegian hospitals with resolutions from $720x576$ to $1920x1072$ pixels and encoded using JPEG compression. The encoding settings can vary across the dataset and they reflect the a priori unknown endoscopic equipment settings. The extension of the image files is ".jpg".

The feature archive contains the extracted visual feature descriptors for all the images from the images archive. The extracted visual features are stored in the text files placed in separate folders and files are named according to the name and the path of the corresponding image files. The extracted visual features are the global image features, namely: Joint Composite Descriptor (JCD) [149], Tamura [135], MPEG-7 [35] features (ColorLayout and EdgeHistogram), Auto Color Correlogram [65] and Pyramid Histogram of Oriented Gradients (PHOG) [42]. Each feature vector consists of a number of floating point values. The size of the vector depends on the feature. The sizes of the feature vectors are: 168 (JCD), 18 (Tamura), 33 (ColorLayout), 80 (EdgeHistogram), 256 (AutoColorCorrelogram) and 630 (PHOG) floating point numbers. Each feature file consists of eight lines, one line per feature. Each line consists of a feature name separated by the feature vector by a colon. Each feature vector consists of a corresponding number of floating point values separated by commas. The extension of each extracted visual feature file is ".features".

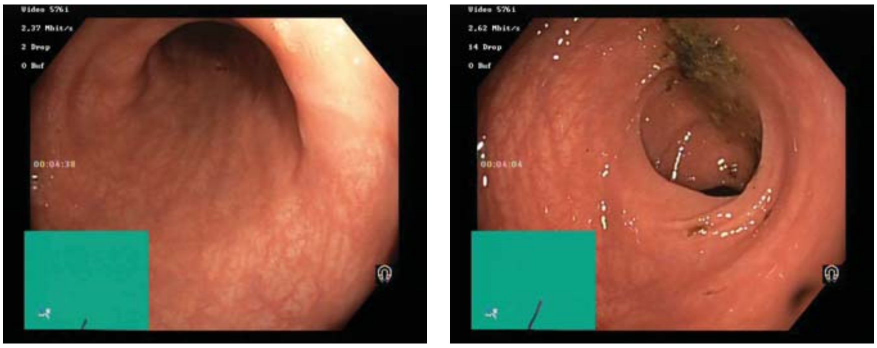
In total, the Medico dataset includes 16 classes showing anatomical landmarks, phatological findings or endoscopic procedures in the GI tract. The anatomical landmarks are Z-line, pylorus and cecum, while the pathological findings include esophagitis, polyps and ulcerative colitis. In addition, we provide two set of images related to the removal of polyps, the "dyed and lifted polyp" and the "dyed resection margins". The dataset includes parts of the Kvasir and Nerthus datasets, but also adds new classes of findings.

Clear Colon

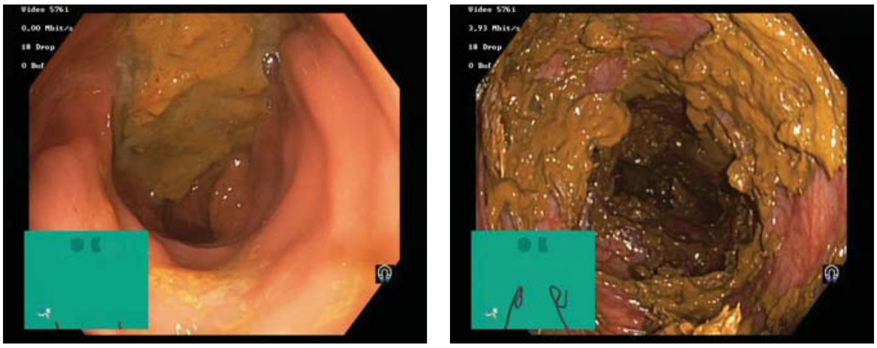
This class represents the samples of normal tissue that can be observed during colonoscopies (see figure 3.5(a) for an example). Comparing to abnormalities, there is no interest in detecting this type of image during live colonoscopies. However, we think that this class can be used for the opposite detection task when the detection algorithm can signal in case of detecting anything that is not normal. This can with a proper implementation and training potentially increase the accuracy for the detection of all other classes.

Stool

Stool is the normal content of the GI tract, consisting of fecal masses and food left-overs. Any fecal mass should be removed before performing colonoscopies and, especially, inter-GI surgical procedures. Despite being a common finding, it is important to be able to detect it because this is a direct indicator of the GI tract preparation quality, which matters for endoscopic procedures' effectiveness. Detected stool masses, even in small pieces, can be considered as a compromising factor to the prior GI tract preparation quality. They can hide small appearances of a very dangerous lesions, e.g. polyps potentially developing into cancer and colon wall penetrations, making stool detection an important task. Moreover, the quality of bowel preparation is considered a key quality indicator for colonoscopy, while directly affecting adenoma detection and decisions on screening and follow-up intervals. Thus, an objective and



(a) Clear colon, no stool masses (b) Inclusions of stool



(c) Medium amount of stool (d) Significant amount of stool

Figure 3.5: The example images depicting different amount of stool masses in the colon.

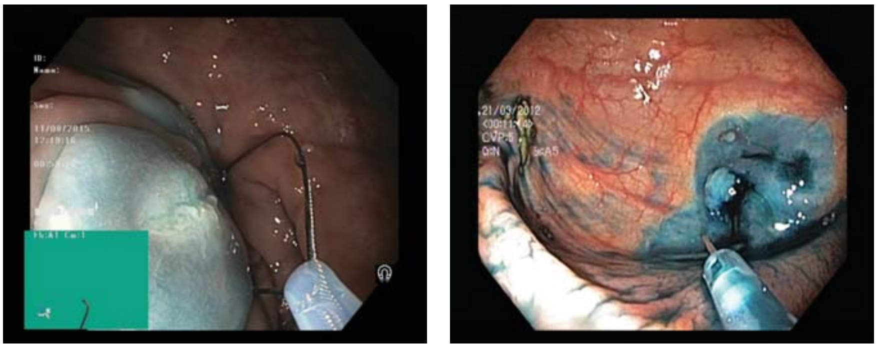
accurate interpretation of the bowel cleanliness is important and, therefore, we added a two new classes, both containing different amounts of stool masses (see figure 3.5 for the examples of stool inclusions 3.5(b), medium 3.5(c) and significant 3.5(d) amounts of stool).

Instruments

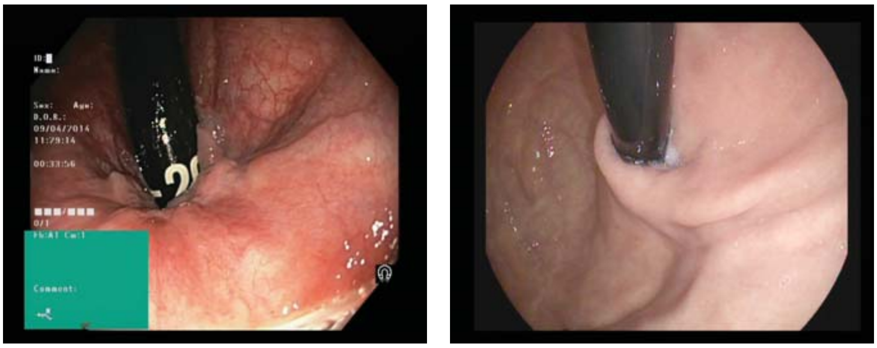
Instruments are artificial objects that can normally be observed in the GI tract during endoscopic procedures. They can be separate auxiliary tools, e.g. expansion nets, balloons, etc., as well as special surgical devices used for interventions and procedures inside GI tract. The detection of instruments during live endoscopies is not a vital task, however it is important to support the reporting process and for the a posteriori analysis of captured data and procedure quality assessment. Moreover, instrument detection and recognition is important for the annotation of the available anonymized datasets. Therefore, in the Medico dataset we introduced three new classes: one depicts different samples of instruments and two others show so-called retroflex vision images. Retroflexing is a special procedure used to get an observation of tissue that is hidden from the doctor's eye during straight-forward endoscope movement. Apart of tissue surface analysis, information extracted from this type of frames can be used as an auxiliary input for precise endoscopy and lesion position localization using the distance marks found on the endoscope's tube. Figure 3.6 depicts examples of the instruments in the Medico dataset.

Auxiliary classes

We also added two auxiliary classes represent images that are useless for lesion detection, but are often appear in non-filtered data captured during routine procedures: blurry frames without any significant content and out-of-patient images that are captured before or after an



(a) Endoscopic surgical snare (b) Endoscopic syringe



(c) Retroflex of rectum (d) Retroflex of stomach

Figure 3.6: Images depicting various instruments including manipulating devices (a) and (b), and endoscope itself captured via retroflex action (c) and (d).

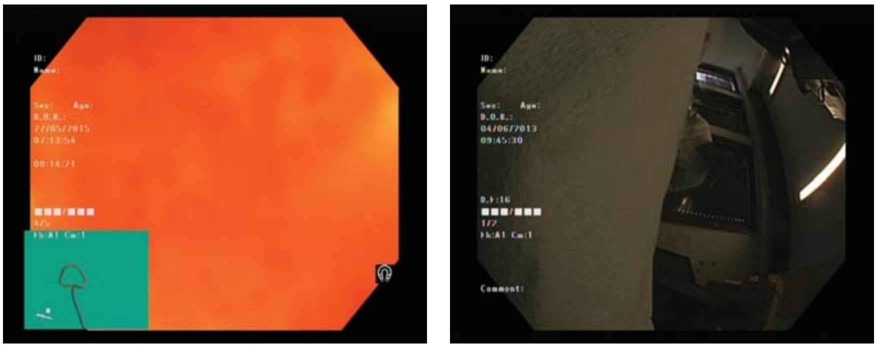
endoscopic procedure (see figure 3.7). Out-of-patient images can also be used for detection the begin and end of endoscopic procedure, which is important for automated reporting generation.

3.1.3.4 Further Dataset Development

The Kvasir, Nerthus and Medico datasets became quite popular open datasets for the research community. We plan to further improve the quality and the size of the datasets by adding new classes of findings, introducing detailed ground truth masks showing the exact location of the findings in each frame, and extending the datasets with VCE-captured frames and videos. The upcoming important classes include colorectal cancer, GI tract bleeding and angioectasia lesion.

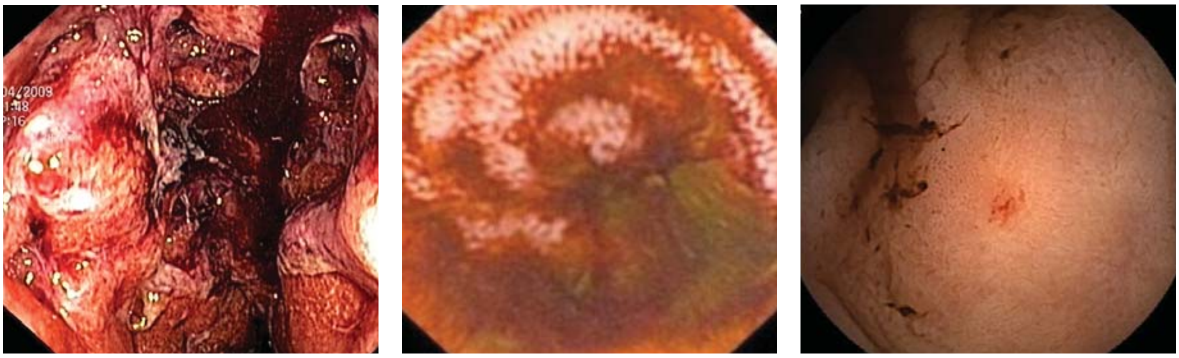
Colorectal Cancer

CRC is the development of cancer from the colon or rectum, which are parts of the large intestine. In the same way as other types of cancer, CRC is the abnormal growth of cells that have the ability to invade or spread to other parts of the body. CRC is a major health issue world-wide. It has one of the highest incidences and mortality of the diseases in the GI tract (see figure 3.8(a) for an example), early detection is essential for a good prognosis and treatment [116]. Several screening methods for CRC exist, e.g., fecal immunochemical tests (FITs), sigmoidoscopy screening, computed tomography (CT) scans and, the most reliable one, traditional colonoscopy.



(a) Blurry frame (b) Out-of-patient

Figure 3.7: Images depicting auxiliary image classes: (a) blurry frames without any recognizable content, and (b) out of the patient images.



(a) Colorectal Cancer (colonoscopy) (b) GI Bleeding (VCE) (c) Angioectasia (VCE)

Figure 3.8: Images depicting various classes will be added to our open datasets in the near future

Gastrointestinal Bleedings

Gastrointestinal bleeding, also known as gastrointestinal hemorrhage, covers all forms of bleeding in the GI tract. It can range from small and hard-to-notice spots without any symptoms to significant blood loss over a short time (see figure 3.8(b) for an example), including symptoms like vomiting red blood, vomiting black blood, bloody stool, or black stool. The bleeding is mostly caused by severe gastric diseases like infections, cancers, vascular disorders, adverse effects of medications, and blood clotting disorders. Common diagnostic procedures include stool sampling, fecal bio-markers analysis, traditional push and modern VCE endoscopy.

Angiectasia

Angiectasia, formerly called angiodysplasia, is one of the most frequent vascular lesions. It is a small vascular malformation of the gastrointestinal wall (see figure 3.8(c) for an example). It is a common cause of otherwise unexplained gastrointestinal bleeding and anemia, and often a source of gastrointestinal bleedings. Lesions are often occur in groups, and they do frequently involve the cecum or ascending colon, although they can occur at other places. The diagnosis of angiectasia is usually performed with push enteroscopy. The lesions can be notoriously hard to find and can be located in hard-to-reach regions of GI tract, eg. the small bowel.

3.1.3.5 Application of the Datasets

Our vision is that the available data may eventually help researchers to develop systems that improve the health-care system in the context of disease detection in videos of the GI tract. Such a system may automate video analysis and endoscopic finding detection in the esophagus, stomach, bowel and rectum. Important results include higher detection accuracies, reduced manual labor for medical personnel, reduced average cost, less patient discomfort and possibly an increased willingness to undertake the examination. With respect to the direct use in multimedia research, the main application area of Kvasir is automatic detection, classification and localization of endoscopic pathological findings in an image captured in the GI tract. Thus, the provided dataset can be used in several scenarios where the aim is to develop and evaluate algorithmic analysis of images. Using the same collection of data, researchers can compare approaches and experimental results directly, and results can easily be reproduced. In particular in the area of image retrieval and object detection, Kvasir will play an important initial role, where the image collection can be divided into training and test sets for the development of various image retrieval and object localization methods including search-based systems, neural-networks, video analysis, information retrieval, machine learning, object detection, deep learning, computer vision, data fusion and big data processing.

Our vision is that the available data may eventually help researchers to develop systems that improve health-care in the context of the GI tract endoscopic diagnosis. Adequate bowel preparation (cleansing) is required to achieve high quality colonoscopy examinations. We invite multimedia researchers to contribute to the medical field by making systems that automatically and consistently can evaluate the quality of bowel cleansing. Innovations in this area that contribute computer-aided assessment and automatic reporting may potentially improve the medical field of GI endoscopy. In the end, the improved quality of GI tract investigations will probably significantly reduce mortality and the number of luminal GI disease incidents.

3.2 Data Exploration, Annotation and Visualization Subsystem

User-guided interactive exploration of big image collections is an important task in many scientific and applied domains. Examples include medical, satellite and industrial image analysis, security, social media and news analysis, and personal photos. Despite the many new and powerful automated image analysis and clustering software, the human eye remains the most important analytic instrument. Research on the topic of interactive image database visualization [103] confirms the importance of human-accessible representation in combination with image clustering, annotation and tagging. Existing image processing tools and frameworks demonstrate interesting and promising approaches, and they give wide opportunities for image browsing, content analysis and performing various data analytic tasks. However, there is still a lack of tools that implement both fast and efficient image collection visualization together with image content analysis and annotation. Moreover, in the medical field, the amount of time experts can use for data annotation is quite limited. This is primarily because of high every-day workloads for doctors. Even further, the annotation of images and videos itself is very time-consuming, and the quality of annotations depends on the experience and concentration of the doctors [53]. For example, in a VCE procedure, a video containing around $216,000$ - $1,000,000$ frames per examination is produced. An experienced endoscopist usually needs from one to two hours to only view and analyze all the video data without performing detailed annotation [76]. Therefore, we developed the automated data exploration, visualization and annotation subsystem is able to reduce annotation workload.

Our approach to efficient data exploration and annotation is based on content-based image retrieval [43] and utilizes number of different techniques and methods for interactive visualization and clustering for unsupervised knowledge discovery in the various image analysis domains providing the outstanding visualization performance for vast collections of images. The developed software made as the universal solution and it is usable not only for medical, but for any use case that involves interactive browsing, visual analysis and annotation of a large amount of image or video data.

3.2.1 Hyperbolic-Tree-Based Visualization and Clustering

Our software for complex image collection analysis is an explorative hyperbolic-tree-based clustering tool for unsupervised knowledge discovery. The software implements a complete prototype of five-stage information visualization including:

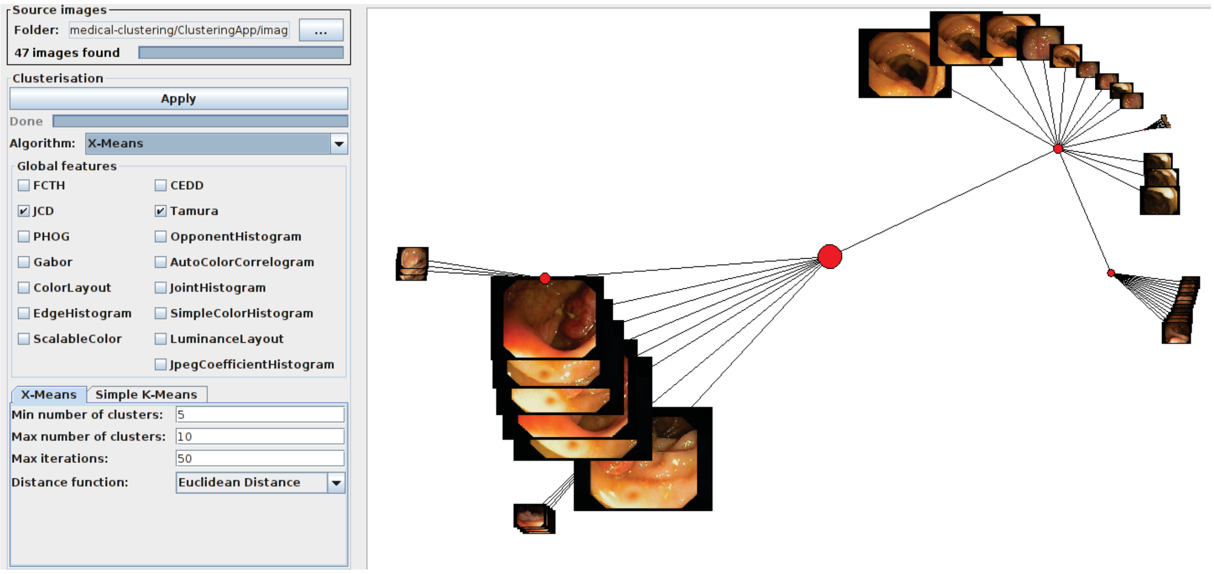
- Raw image and video frames data indexing and loading.
- Analytical abstraction generation via image feature descriptors.
- Visualization abstraction generation via clustering, centroids and distance values computation.
- User-view generation via interactive hyperbolic tree.
- Metadata generation during interactive clusters exploration.

The software is written in Java and uses two open-source libraries, LIRE and WEKA[1] [58] for image features extraction and clusterization. LIRE is a library that supports multiple global and local image features out of the box. Here we use Color and Edge Directivity Descriptor (CEDD) [37], Joint Composite Descriptor (JCD) [149], Fuzzy Color and Texture Histogram (FCTH) [38], Tamura [135], Pyramid Histogram of Oriented Gradients (PHOG) [42], Auto Color Correlogram [65], Local Binary Patterns [57], and MPEG-7 [35] features including Edge Histogram, Color Layout and Scalable Color. WEKA is a collection of tools for machine learning and data mining. It can be directly combined with the LIRE code for easy integration. Here we use X-means, K-means and hierarchical clustering algorithms.

Initially, the prototype was designed as an interactive demo with a live and responsive view that allowed users to interact with the created clusters and their hyper-tree representation. Clustering performed using image features and folder structure if desired. We used two datasets: one with still pictures showing disease symptoms in a medical scenario, another with pictures of the

---

[1] `http://www.cs.waikato.ac.nz/ml/weka/`



Folder: medical-clustering/ClusteringApp/imag

47 images found

Algorithm: X-Means

☐ FCTH
☐ CEDD
☑ JCD
☑ Tamura
☐ PHOG
☐ OpponentHistogram
☐ Gabor
☐ AutoColorCorrelogram
☐ ColorLayout
☐ JointHistogram
☐ EdgeHistogram
☐ SimpleColorHistogram
☐ ScalableColor
☐ LuminanceLayout
☐ JpegCoefficientHistogram

Min number of clusters: 5
Max number of clusters: 10
Max iterations: 50
Distance function: Euclidean Distance

Figure 3.9: Hyper-tree based visualization, clustering and annotation system.

same tagging categories in a social image collection. Despite the initial demo-development purposes, our prototype showed great potential and not restricted to a specific domain.

Figure 3.9 shows a screen shot of the demo application. Users can interactively perform the following operations:

- Select the folder containing the image collection.
- Select Clustering algorithm and its parameters.
- Choose one or several different image features. If more than one feature is picked, they will be combined using early fusion.
- Initiate features extraction and clusterization process.
- Interact with the hyper-tree by zooming and turning it into different angles.
- Inspect cluster and individual image properties and name/tag the images and clusters.

Practical usage experience by domain experts who used this hyperbolic-tree-based visualization approach confirmed the importance of the unsupervised clustering algorithms to explore image and video data collections that do not contain meta-data. Our clustering methodology leads to good annotation results, and therefore, provides a good method for the abstraction stages. However, as a result of a successful collaboration with Norwegian hospitals, we have collected a large dataset consisting of more than 77.000 images and 600 videos from medical procedures. The size of this unannotated data collection was too big for efficient processing with this first application due to memory constrains and drawing performance issues. Thus, we continued our development focusing on support for handling big data collections.

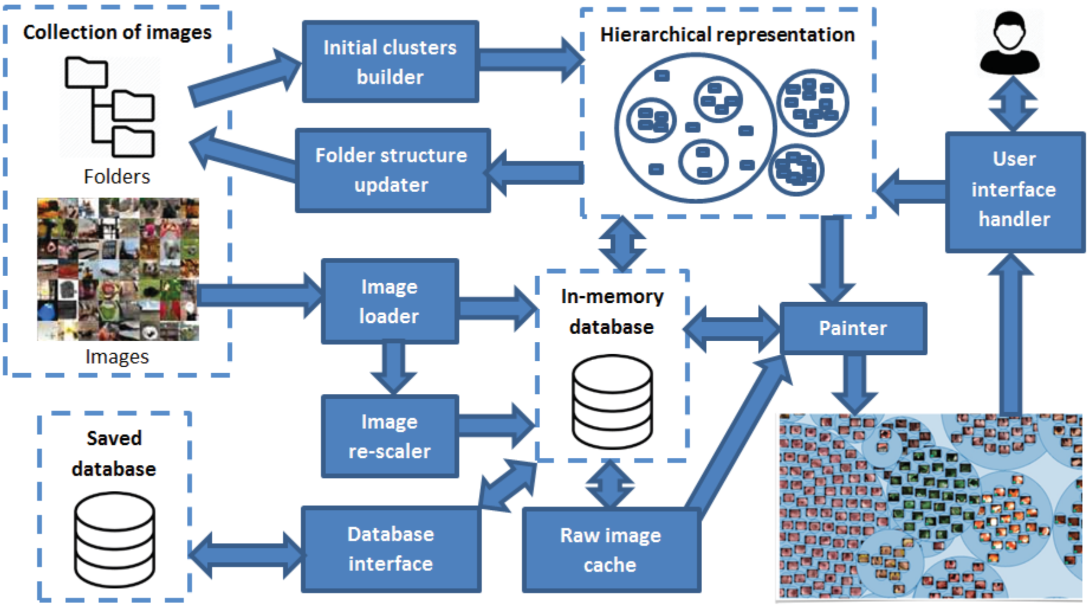


Figure 3.10: Structure of the visualization and user interface engine of the presented ClusterTag application. A number of caching and intermediate data processing routines are used to make it possible to perform real-time visualization and interaction with huge image collections.

3.2.2 Cluster-Based Visualization and Annotation (ClusterTag)

To solve the visualization performance issues that we met during the use of the hyper-tree-based visualization and provide more efficient solution for visualization and annotation, we performed a software structure redesign involving a set of modifications and improvements to extend the tool to make it universal and usable for any use case involving interactive browsing, visual analysis and the annotation of a large amounts of image or video data. As a result, the redesigned software named ClusterTag [98] does now have the following advanced properties:

- It allows users to investigate and analyze vast collections of images by providing a configurable focus and context view based on similarity of frames.
- It provides a focus and context view for annotation and tagging of the dataset, making it more accessible for complementary information systems.
- The tool structure is flexible and it can be easily adapted to different use cases and extended with new image processing algorithms.
- It supports real-time, interactive viewing, analysis and modifications of the dataset, giving new opportunities for on-line-like data analytics.

One of the main features of our ClusterTag application is interactivity with a visual collection representation. Users interact with the images and the created or already defined clusters. In this application, we use LIRE and WEKA for image features and clustering support, respectively. Additionally, ClusterTag is build in a modular way allowing for easy replacement of WEKA and LIRE by other machine-learning or feature-extraction libraries if desired.

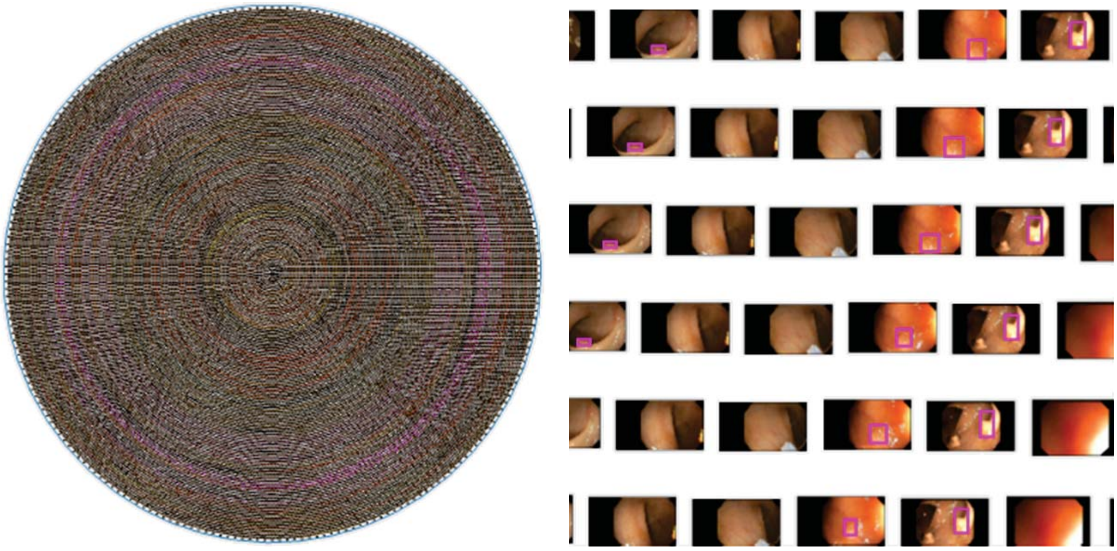
To be able to implement a visualization tool for a virtually unlimited number of images simultaneously in real-time and give the user the ability to interact with them, we developed an optimized visualization engine written in Java. The overall structure of the software is depicted in figure 3.10. It consist of the following sub-modules:

- The initial analyzer of the image collection file and folder structure. The initial folder structure is used to form the initial image clusters.
- The painter module used to draw the user interface and the visual representations of the cluster hierarchical structure.
- The image-oriented in-memory database and the image cache, implementing the optimized image preloading, rescaling and drawing.
- Off-line on-disk mirror copy updater and annotation meta-data saver responsible for updating the collection's file structure on the disk after any modification done to the clusters by the user or by the clustering procedure.

We have designed and implemented several additional optimization techniques to allow real-time handling of huge image collections. The most important are a database of ready-to-draw pre-processed images, caching of raw image visual representation, painting of adaptive image spatial resolution, interaction with partially processed collections, multi-scale image painting, multi-threaded image processing and feature extraction (see figure 3.10 for an overview). Even further, to speed-up and smoothe the annotation process, we provide the ability to start exploring the image collection immediately regardless of the image pre-processing and feature extraction progress. In case of a newly opened collection, a visual representation becomes available immediately after the initial directory structure listing and the visual representation is updated in correspondence with the collection processing progress.

The ClusterTag application, first, allows users to choose the folder containing the image collection. Immediately after listing the files of a new image collection, it appears in the main window as it was organized in the folder structure, and the user can immediately start exploring the collection. Figure 3.11(a) shows a visualization of an unsorted collection of $36,476$ medical images. The user can navigate through the collection's view using the mouse to move, zoom into and zoom out of the field of view (see figure 3.11(b)). To perform clustering, the user can select a desired clustering algorithm, its parameters and several different image features. If more than one feature is selected, they will be combined using early fusion. After selecting all the parameters, the user can apply clustering to the dataset creating the clusters. Figure 3.11(c) shows a visualization of the collection of medical images clustered using the JCD and Tamura global image features, which produces a number of dense clusters representing visually similar images in the same clusters. The zoomed view of the clustered collection is depicted in figure 3.11(d). The cluster leaves are represented using the image that is closest to the cluster center, i.e., the cluster medoid. Individual images and image groups can be dragged and dropped between different clusters reflecting changes to the file structure of the collection. It is possible to name/tag the clusters and individual images.

The ClusterTag tool was intensively used during the Kvasir and Nerthus dataset creation and annotation. It already demonstrated a great potential for big image processing and was evaluated with different end-users and domain experts including experienced medical doctors.


(a) Unsorted collection of images. (b) Zoomed view of the unsorted collection.

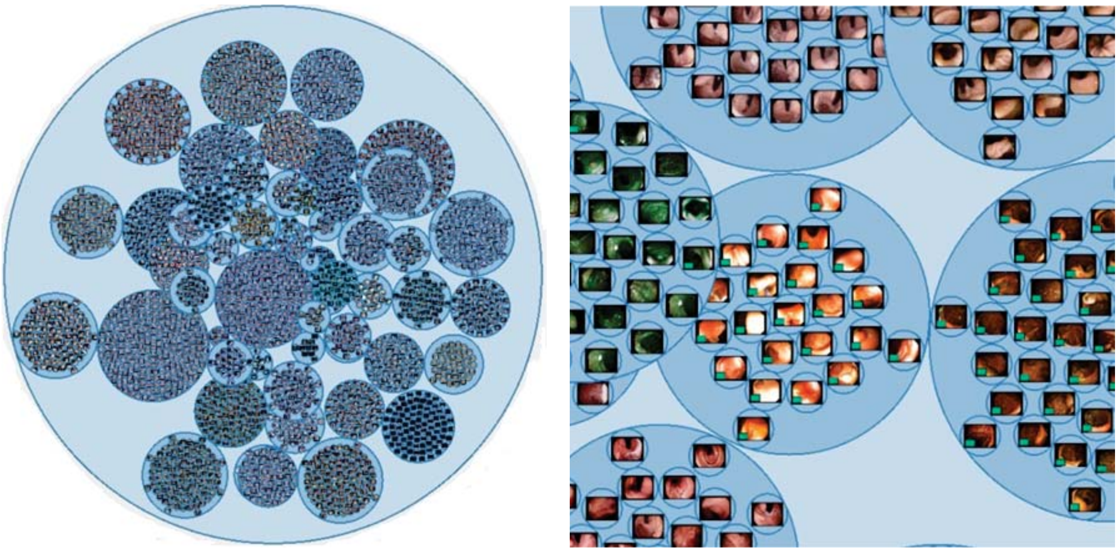

(c) Clustered collection of images. (d) Zoomed view of the clustered collection.

Figure 3.11: Examples of visual representations of an image collection containing $36,476$ unsorted medical images generated by the ClusterTag application. The initial view of the loaded collection shows all the images in one big cluster. After the clustering, using the JCD and Tamura global image features, the software generates a number of dense clusters representing visually similar images in the same clusters.

3.3 Detection Subsystem

The detection subsystem performs lesion or object recognition and classification. It is intended for abnormality- or object-presence detection without searching for their precise position. The detection is performed using various computer vision, visual similarity finding, deep- and machine-learning-based techniques. For each lesion that has to be detected, we use a set of reference frames that contains examples of this lesion occurring in different parts of the GI tract. This set can be seen as the model of the specific disease. We also use sets of frames containing examples of all kinds of healthy tissue, normal findings like stool, food, liquids, etc.

The final goals of the detection subsystem is to decide if this a particular analyzed frame contains any lesion (detectable object) or not, and to detect the exact type of the lesion (detectable object). The detection system is designed in a modular way and can easily be extended with new diseases. This would, for example, allow not only to detect a polyp, but to distinguish between a polyp with low or high risk for developing CRC [96].

3.3.1 Single-Class Global-Feature-based Detection

In our previous work [112], we presented our basic EIR system [101, 117, 118] that implements a single-class global-feature-based detector able to recognize the abnormalities in a given video frame. Global image features were chosen, because they are easy and fast to calculate [79], and the exact lesion's position is not needed for detection, i.e., identifying frames that contain a disease. We showed [97] that the global features we chose [115] can indeed outperform or at least reach the same results as local features [112].

The basic algorithm is based on an improved version of a search-based method for image classification. The overall structure and the data flow in the basic EIR system is depicted in figure 3.12. First, we create the index containing the visual features extracted from the training images and videos, which can be seen as a model of the diseases and normal tissue. The index also contains information about the presence and type of the disease in the particular frame. The resulting size of the index is determined by the feature vector sizes and the number of required training samples, which is rather low compared to other methods. Thus, the size of the index is relatively small compared to the size of the training data, and it can easily fit into the main memory on a modern computer. Next, during the classification stage, a classifier performs a search of the index for the frames that are visually most similar to a given input frame (see section 3.3.2 for a detailed description of the method). The whole basic detector is implemented as two separate tools, an indexer and a classifier. We have released the indexer and the classifier as an open-source project called *OpenSea*[2] [90].

The indexer is implemented as a batch-processing tool. Creating the models for the classifier does not influence the real-time capability of the system and can be done off-line, because it is only done once when the training data is first inserted into the system. Visual features to calculate and store in the indexes are chosen based on the type of the disease because different sets of features or combinations of features are suitable for different types of diseases. For example, bleeding is easier to detect using color features, whereas polyps require shape and texture information.

The classifier can be used to classify video frames from an input video into as many classes as the detection subsystem model consists of. The classifier uses indexes generated by the indexer. In contrast to other classifiers that are commonly used, this classifier is not trained in a separate learning step. Instead, the classifier searches previously generated indexes, which can be seen as the model, for similar visual features. The output is weighted based on the ranked list of the search results. Based on this, a decision is made. The classifier is parallelized and can utilize multiple CPU cores for the extraction of features and the searching in indexes. To increase performance even more, we implemented the most compute intensive parts of the system with GPU computation support.

---

[2] `https://bitbucket.org/mpg_projects/opensea`

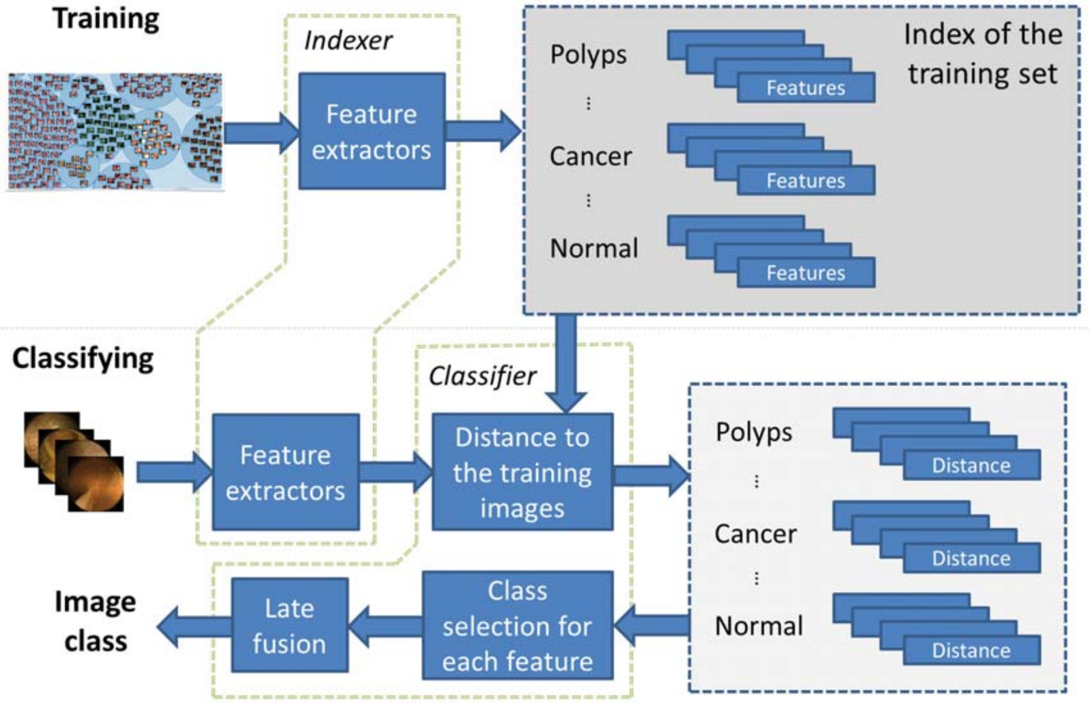


Figure 3.12: Detailed steps for the multi-class global-feature-based detection implementation

3.3.2 Multi-Class Global-Feature-based Detection

The multi-class global-feature-based detector is based on our search-based classification algorithm that is used to create a classifier for each disease that we want to classify. Figure 3.12 gives a detailed overview of the classifier's pipeline for the global-feature-based implementation of the detection. The difference to the basic EIR version is that the ranked lists of each search-based classifier are then used in an additional classification step to determine the final class.

For feature extraction in the detection step and for the training procedure, the indexing is performed using the basic EIR indexer implementation [101, 118]. The same set of two global features, namely Tamura and JCD, is used. These features were selected using a simple feature efficiency estimation by testing different combinations of features on smaller reference datasets to find the best combinations in terms of processing speed and classification accuracy. The selected features can be combined in two ways. The first is called feature value fusion or early fusion, and it basically combines the feature value vectors of the different features into a single representation before they are used in a decision-making step. The second one is called decision fusion or late fusion and the features are combined after a decision-making step. Our multi-class global-feature-based approach implements feature combination using the late fusion.

During the detection step, a term-based query from the hashed feature values of the query image is created for each image, and a comparison with all images in the index is performed, resulting in a ranked list of similar images. The ranked list is sorted by a distance or dissimilarity function associated with the low-level features. This is done by computing the distance between the query image and all images in the index. The distance function for our ranking is the Tanimoto distance [136]. A smaller distance between an image in the index and the query image means a better rank [136]. The final ranked list is used in the classification step, which implements a simple k-nearest neighbors algorithm [19]. This algorithm can be used for supervised and unsupervised learning, two or multi-class classification and different types of input data ranging from features extracted from images to videos to meta-data. Its main advantages are its simplicity, that it achieves state-of-the-art classification results and that it is computationally very cheap.

For the final classification, we use the random forest classifier [29], an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes of the individual trees. A decision tree can be seen as a classifier, which basically performs decision-based classification on the given data. To get the final class, the classifier combines decision trees into a final decision implementing a late fusion for the multi-class classification. The advantage of the random forest algorithm is that the training of the classifier is very fast because the classification steps can be parallelized since each tree is processed separately. Additionally, it is shown [142] that the random forest is very efficient for large datasets due to the ability to find distinctive classes in the dataset and also to detect the correlation between these classes. The disadvantage is that the training time increases linearly with the number of trees. However, this is not a problem for our use-case since the training is done offline, where time is less critical. Our implementation of the random forest classifier uses the version provided by WEKA. It is important to point out that for this step, another classification algorithm can also be used.

3.3.3 Deep-Learning-based Detection

The neural network version of EIR called Deep-EIR is based on a pre-trained convolutional neural network architecture and transfer learning [33]. We trained a model based on the InceptionV3 architecture [132] using the ImageNet dataset [44] and then re-trained and fine-tuned the last layers. We did not perform complex data augmentation at this point and only relied on transfer learning for now. For future work, we will also look into data augmentation and training a network from scratch using the newly collected data, which might lead to better results than transfer learning.

Figure 3.13 gives a detailed overview of the complete pipeline for the neural network-based implementation of the detection based on multi-class image classification.

InceptionV3 achieves good results regarding single-frame classification and has reasonable computational resource consumption. It is built on top of Google's Tensorflow [12], which provide a framework for numerical computations using graphs, especially neural network-based architectures. We used a pre-trained InceptionV3 model [132] with the following retraining step. For retraining, we follow the approach presented in [46]. Basically, we froze all the basic convolutional layers of the network and only retrained the two top fully connected layers. The fully connected layers were retrained using the RMSprop [139] optimizer that allows an adaptive learning rate during the training process. After 1,000 epochs, we stopped the retraining of the FC layers and started fine-tuning the two top convolutional layers. This step finalizes the transfer-learning scenario and performs an additional tuning of all the NNs layers according to our dataset. For this training step, we used a stochastic gradient descent method with a low learning rate of 10^{-4} to achieve the best effect in terms of speed and accuracy [85]. This comes with the advantage that little training data is needed to train the network, which is an advantage for our medical use case. Additionally, it is fast, requiring just about one day to retrain the model. Our re-trainer is based on an open-source implementation[3]. To increase the number of training samples and reduce overfitting of the model, we also performed distortion operations on the images. Specifically, we performed random cropping, random rescaling and random change of brightness. The grade of distortion was set to 25% per image. After the model has been retrained, we use it for a multi-class classifier that provides the top five classes based on probability for each class.

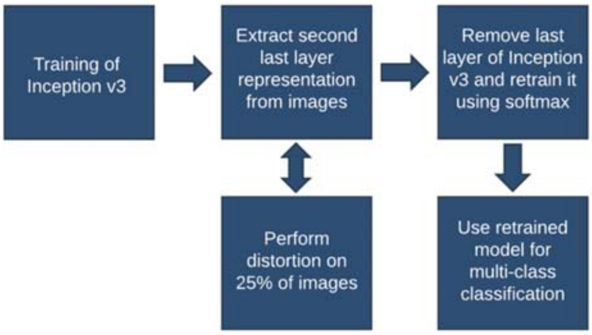

Training of Inception v3
Extract second last layer representation from images
Remove last layer of Inception v3 and retrain it using softmax
Perform distortion on 25% of images
Use retrained model for multi-class classification

Figure 3.13: Multi-class deep-learning-based detection pipeline

3.3.4 Deep-Feature-based Detection

Our deep-feature-based detection (see figure 3.14) approach is designed using different well-known-working deep learning architectures to extract either the features directly or to classify the images using the whole range of concepts and their probabilities as input for the various machine-learning-based classifiers. The architectures that we used are ResNet50 [59], VGG19 [125], InceptionV3 [133] and Xception [40].

Here, we use only pre-trained models of the mentioned architectures in two main modes: deep-feature and deep-concept extraction. Deep feature is the vector of floating point numbers that represents an output of the pre-top-layer of the deep convolutional neural network (DCNN) architecture. Normally, this vector is used as an input to the top fully connected layers of the DCNN, thus it represents the highest-possible vector of the image features used for the final image classification on the top layers. In case of already pre-trained an DCNN, the deep feature vector contain information about all the image's high-level features in a compact form. For the used architectures, the size of the vector with deep features is pre-defined [93] and it does neither depend on its single- or multi-class nature, nor on the number of classes supported by the specific DCNN. In contrast, deep concept is an output of the top layer of the multi-class classification DCNN. That is a vector of floating point numbers with the size equal to the number of classes for that this particular DCNN. The deep concept vector represents the detection probabilities of the each and every DCNN-supported concept. Here, the meaning of concept is equal to the meaning of class for multi-class classification problems. The main difference is that in our approach the concepts' probabilities are not the final output of the detector, but they are used as a feature vector in the further stages of detection.

---

[3] `https://github.com/eldor4do/Tensorflow-Examples/blob/master/retraining-example.py`

The DCNN models are used as is without any additional retraining, and we rely on the transfer learning methodology for the final detection. After extracting the corresponding deep features or/and concepts, they are used as the input to the classical machine-learning-based multi-class classifiers. We use Random Tree [28], Random Forest [29] and Logistic Model Tree [130] classifiers that were proven to perform efficiently and are able to process the feature vectors at a reasonable speed.

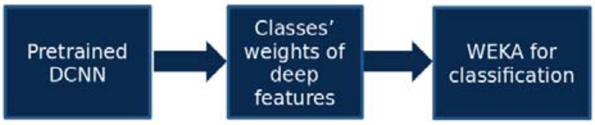


Figure 3.14: DCNN concepts- and deep-features-based detection pipeline

3.4 Localization Subsystem

The localization subsystem is intended for finding the exact positioning of a lesion, which is used to show markers or areas in the frame containing the disease. This information is then used by the visualization subsystem. The localization subsystems can be used in combination with multiple analytic modules designed for various diseases and different localization precision. All modules are divided into two main classes depending on the input data requirements: position finders and complete localizers. The position finders require preliminary frames' processing by the corresponding detection subsystem and process only frames marked as positive by the detection subsystem. Complete localizers provide the integral solution to the disease finding problem. First, they process the whole frame and perform its fine or/and coarse segmentation with box- or pixel-wise granularity. Then, this segmentation information is used for both exact lesion position marking and disease presence detection. Therefore the complete localizers do not require preliminary frames' processing by the corresponding detection subsystem and, despite they higher complexity, can even perform faster in terms of the overall detection plus localization performance.

3.4.1 Hand-Crafted Local-Feature-based Position Finder

The local-feature-based position finder is designed as a pipelined frame processor that utilizes several hand-crafted local image features in order to perform localization of polyps. Processing is implemented as a sequence of intra-frame pre- and main-filters. Pre-filtering is required because we use local image features to find the exact position of objects in the frames, and these features can be affected by pixel noise and local color defects. In general, lesion objects or areas can have different shapes, textures, colors and orientations. They can be located anywhere in the frame and can also be partially hidden and covered by biological substances, like seeds or stool, and lighted by direct light. The image itself can be interlaced, noisy, blurry and over- or under-exposed, and it can contain borders and sub-images. Images can have various resolutions depending on the type of endoscopy equipment used. Endoscopic images usually have a lot of flares and flashes caused by a light source located close to the camera. All these nuances

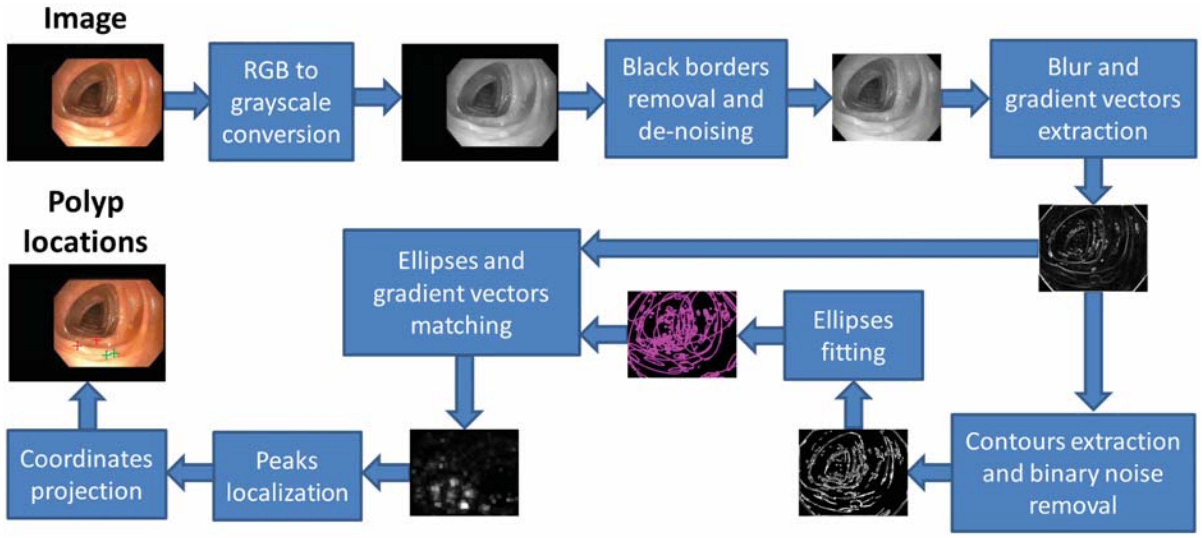


Image
RGB to grayscale conversion
Black borders removal and de-noising
Blur and gradient vectors extraction
Polyp locations
Ellipses and gradient vectors matching
Ellipses fitting
Contours extraction and binary noise removal
Coordinates projection
Peaks localization

Figure 3.15: Detailed steps of the hand-crafted local-feature-based localization algorithm implementation

affect the local feature-based localization methods negatively and have to be specially treated to reduce localization precision impact. In our case, several sequentially applied filters are used to prepare raw input images for the following analysis. These filters are border and sub-image removal, flare masking and low-pass filtering. After pre-filtering, the images are ready to be used for further analysis.

The main localization algorithm able to spot colon polyps using our hand-crafted approach is based on several local image features [116]. The main idea of the localization algorithm is to use the polyp's physical shape to find the exact position in the frame. In most cases, the polyps have the shape of a hill located on a relatively flat underlying surface or the shape of a more or less round rock connected to an underlying surface with stalks of varying thickness. These polyps can be approximated with an elliptically shaped region consist of local features that differ from the surrounding tissue. To detect these types of objects, we process frames marked by the detection subsystem as containing polyps by a sequence of various image processing procedures, resulting in a set of possible abnormality coordinates within each frame. Figure 3.15 gives a detailed overview of a localization pipeline. The pipeline consists of the following steps: non-local means de-noising [31]; 2D Gaussian blur and 2D image gradient vector extraction; border extraction by gradient vector threshold binarization; border line isolated binary noise removal; estimation of ellipse locations; ellipse size estimation by analyzing border pixel distribution; ellipse fitting to extracted border pixels; selection of a predefined number of non-overlapping local peaks and outputting their coordinates as possible polyp locations. For the possible locations of ellipses, we use the coordinates of local maxima in the insensitivity image, created by additive drawing of straight lines starting at each border pixel in the direction of its gradient vector. Ellipse fitting is then performed using an ellipse fitting function [49].

3.4.2 Deep-Learning-based Region Localizers

Despite the promising performance shown by the hand-crafted polyp finder, it is limited to polyps and is hard to extend toward other flat lesions or findings that can vary in shape and properties, like, eg., ulcer lesion and Z-line landmark. The next generation of complete localizers used in DeepEIR system are deep-learning-based region localizers. The idea of utilizing deep-learning-based methods for the localization tasks appeared in connection with the need to simplify support for adding different diseases by implementation of lesion-specific shape, color and texture detection, which requires a lot of manual work and experimental studies for each new type of abnormality. In order to reduce the system improvement costs, we performed an evaluation of two universal deep-learning-based object localization approaches that were adapted to fit the processing requirements of medical imaging. The first is TensorBox[4] [128], which extends Google's Tensorflow DCNN reference implementation [12]. The second approach is based on the Darknet [105] open-source deep learning neural network implementation called YOLO[5] [106]. Both of these frameworks are designed to provide not only object detection, but also object localization inside frames.

The TensorBox approach introduces an end-to-end algorithm for detecting objects in images. As input, it accepts images and generates a set of object bounding boxes as output. The main advantage of the algorithm is its capability of avoiding multiple detections of the same object by using a recurrent neural network (RNN) with long short-term memory (LSTM) units together with fine-tuned image features from the implementation of a CNN for visual object classification and detection called GoogLeNet [131].

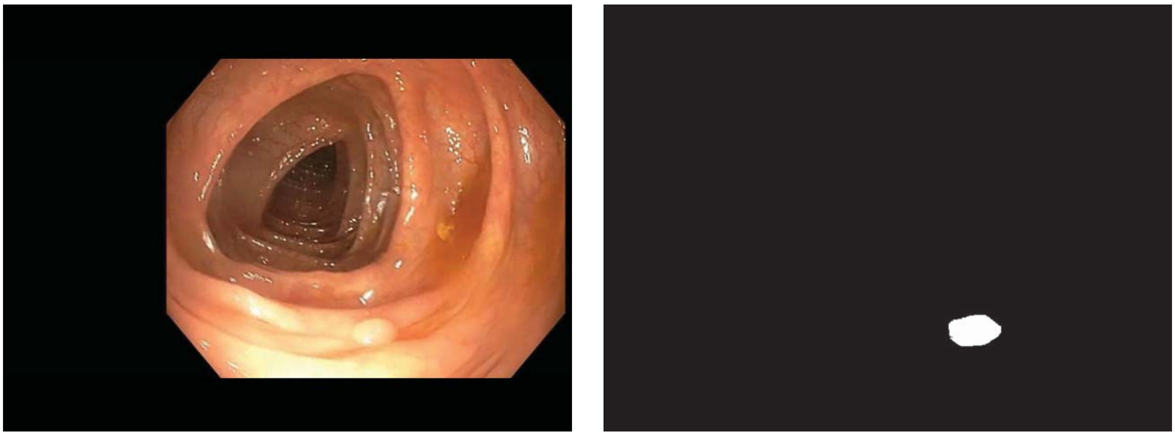
The Darknet-YOLO approach introduces a custom CNN, designed to simultaneously predict multiple bounding boxes and class probabilities for these boxes within each input frame. The main advantage of the algorithm is that the CNN sees the entire image during the training process, so it implicitly encodes contextual information about classes as well as their appearance, resulting in a better generalization of objects' representation. The custom CNN in this approach is also inspired by the GoogLeNet [131] model.

As initial models for both approaches, we used database models pre-trained on ImageNet [68]. Our custom training and testing data for the algorithms consists of frames and corresponding text files describing ground truth data with defined rectangular areas around objects: a JSON file for TensorBox and one text file per frame for Darknet-YOLO. Ground truth data was generated using a binary-masked frame set (example shown in figure 3.16). Both frameworks were trained using the same training dataset, where all frames contained one or more visible polyps. No special filtering or data preprocessing was used, thus the training dataset contained high quality and clearly visible polyp areas as well as blurry, noisy, over-exposed frames and partially visible polyps. The models were trained from scratch using corresponding default-model training settings [106, 128]. After the training, the test dataset was processed by both neural networks in testing mode. As a result, the frameworks output JSON (TensorBox) and plain-text (Darknet-YOLO) files containing sets of rectangles, one set per frame, marking possible polyp locations with corresponding location confidence values. These results have been processed using our localization algorithms.

---

[4] `https://github.com/Russell91/TensorBox`

[5] `https://github.com/pjreddie/darknet`



(a) Frame with polyp
(b) Polyp ground truth

Figure 3.16: Example frames showing polyp and its body ground truth area. This is an example of polyps localization task complexity. Polyp body has the same color, texture properties and light flares as surrounding normal mucosa

3.4.3 Deep-Feature-based Region Localization

The deep-feature-based complete region localization approach is our attempt to utilize our frame-wise deep-feature-based detection algorithm for localization purposes. We have applied the RT-D method to the set of sub-frames generated from the training and test sets. Sub-frames (blocks) are generated using sliding square window with 66% overlap with the neighboring sub-frames. We have tested different window sizes from 64x64 to 128x128 pixels. The best results were obtained using 128x128 windows size. The generated sub-frames are fed into the RT-D detection algorithm, and then, the processed sub-frames are grouped back into the frame. This results in a coarse localization map which is then used for frame-wise detection. The detection is achieved by applying a simple threshold activation function, and we evaluated the activation thresholds ranging from 1 block to 50% of the frame blocks. The best detection results were achieved with a threshold value of 2 blocks.

3.4.4 GAN-based Segmentation, Localization and Detection

The most advanced GAN-based complete segmentation localizer provides a fine pixel-wise marking of the frames with the lesion-occupied areas. It shows not only the location of lesions on the generated segmentation maps, but also provides a probability for each pixel of input image to belong to the lesion area, enabling the efficient and flexible detection-via-localization post-processing of segmentation data. Moreover, this localizer can be easily to various types of lesions regardless of their properties. At the moment we have implemented this localization and the corresponding detection-via-localization for polyps, angiectasia lesion, bleeding and even for non-GI-tract- and non-medical-related objects like spermatozoons, flooded areas, etc.

The proposed segmentation approach (see figure 3.17) is able to mark the object in the given frame with pixel accuracy. To achieve this, we use GAN to perform the segmentation. GANs [54] are machine learning algorithms that are usually used in unsupervised learning and are implemented by using two neural networks competing with each other in a zero-sum game. Modern architectures of GANs have been shown to achieve promising results in terms of per-

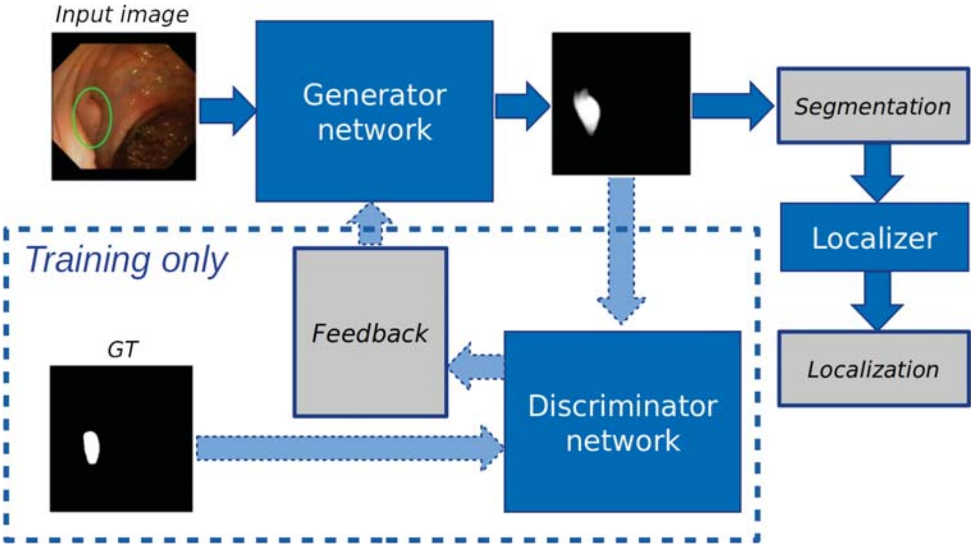
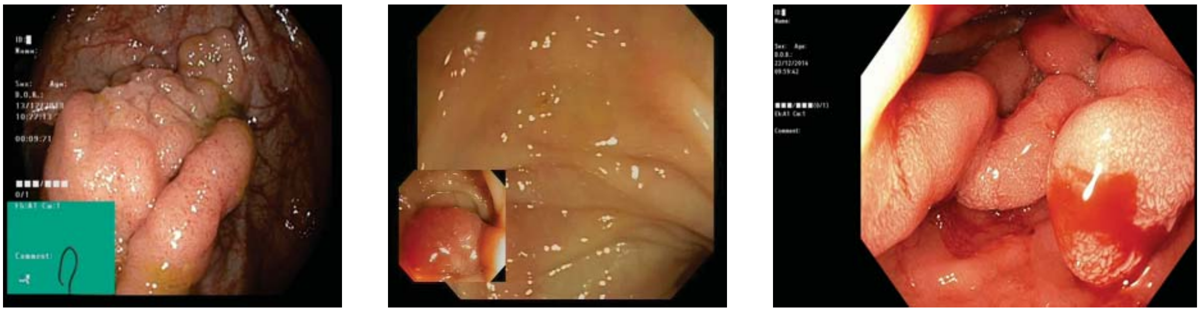


Figure 3.17: GAN-based segmentation and localization pipeline

formance and data processing speed in various image segmentation tasks. They not only can efficiently extract and summarize the local texture and shape properties of the target objects using relatively small training sets, but also can resist the various image property variations, like change of noise level, slight color and luminosity shifts, etc. We use a GAN model initially developed for retinal vessel segmentation in fundoscopic images, called V-GAN. We choose V-GAN as the basis for our polyp segmentation approach development because it demonstrated [127] the good segmentation performance for the retinal images that have the visual properties comparable to the GI tract images. The V-GAN architecture [127] is designed for RGB images and provides a per-pixel image segmentation as output. To be able to use the GAN architecture in our segmentation approach, we added an additional output layer to the generator network that implements an activation layer with a step function that must generate the binary segmentation output. Furthermore, we added support for gray-scale and RGB color space data shapes for the input layers of the generator and discriminator networks including an additional color space conversion step. Gray-scale support was added to be able to use a single value per pixel input in order to reduce the network architecture complexity, to speed up the model training and data processing parts, and also to implement the processing of modern narrow-band images generated by some types of endoscopic devices.

In the same way as all the machine- and deep-learning-based approaches, the proposed localizer requires preliminary training using an appropriate training set consisting of pixel-wise annotated images. The images used in this research are obtained from standard endoscopic equipment and can contain some additional information fields related to the endoscopic procedure. Some types of the field (see Figure 3.18), integrated into resulting frames shown to the doctor and captured by the recording system, can confuse detection and localization approaches, and lead to frame misclassification (green navigation box) or false positive detection (captured frame with polyp). We have implemented a simple frame preparation procedure that consists of three independent steps: black border removal (including patient-related text fields), navigation localizer map masking and captured still frame masking. All the removed and masked regions are excluded from further frame analysis.

Another problem we meet during the development of this advanced localizer is the lack of well-annotated training samples with detailed ground truth masks. To reduce the impact of the limited training sets, we implemented a data augmentation scheme used in the training process of the GAN. The data augmentation scheme implements image rotation in the range of $\pm180^{\circ}$, horizontal and vertical flipping of frames and image insensitivity alteration in the range of $\pm40\%$. These augmentation parameter values were selected during the initial approach development and preliminary evaluation on the reduced training and validation sets.



(a) Navigation (b) Captured frame (c) Patient information

Figure 3.18: Examples of the different auxiliary information fields integrated into recorded frame: a colonoscope navigation localizer (a), a captured still frame (b) and a patient-related information (c). Images taken from CVC-968 [23] and Kvasir [95].

The GAN-based detection-via-localization approach (see figure 3.19) utilizes a simple threshold activation function, which takes the number of positively marked pixels in the frame as input. In the validation experiments performed using different datasets, we evaluated the activation thresholds from one pixel to a quarter of the frame. The best detection results were achieved with a threshold value of 50 pixels [92], which has been used for the detection experiments.

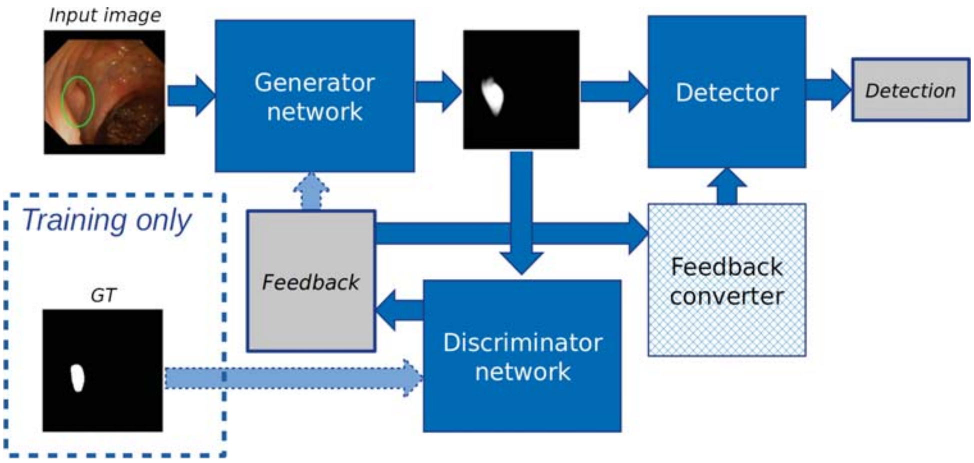


Figure 3.19: GAN-based detection-via-localization pipeline

3.5 Visualization and Results Representation Subsystem

The visualization concepts of the EIR system include multiple different visual data representation strategies. The first-stage data visualization modules were implemented during annotation and visualization subsystem development (see section 3.2). The developed hyperbolic-tree- and cluster-based visualization and clustering approaches demonstrated [120] their great potential for data analysis and were widely used for our own dataset preparation [94, 95, 100]. Further development of the visualization system was necessary for the efficient support of the EIR system user-level task and include both still image (frame) visualization and video stream handling.

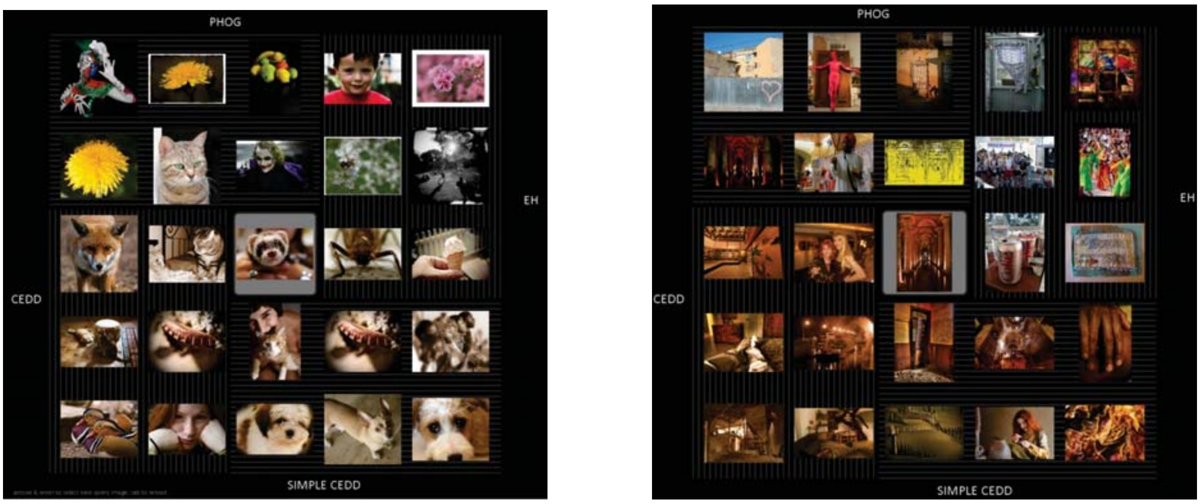
3.5.1 Online Global-Feature-Based Visual Similarity Search Tool

In order to validate our global-feature-based similarity search methodology used in the detection system implementation, we designed and developed an image retrieval and result browsing application, while succeed our previous search-based classifier and visualizer [112]. It utilizes the core strengths of global features: small footprint, high computing and search speed. The tool is unique in its combination of image browsing and searching, where users implicitly select the image features that match their sense of similarity best. At the start, the user provides a query image. Then, the search engine retrieves results using different pre-selected global features. After the users picked the features and used the query image to get the first results, they can explore the available results in four partitions, each representing the results for one feature. Figure 3.20 shows the application's user interface. The query image is shown in the center, lines in the background of the results show the partitions. Users can navigate the search selecting the desired image as the new query image. Therefore, users can browse the data set based on different features. The tool's UI is implemented using the non-commercial open-source version of the QT development library. Feature extraction is implemented via a C++ wrapper for the LIRE library Java API. The tool is cross-platform and can be used from desktop and mobile platforms.

This search and visualization tool allowed us to verify our global-feature-based image matching methodology and demonstrated the validity of the desired approach. The tool was described in [80], presented for the first time at the 7th International Conference on Multimedia Systems, and received positive feedback from the multimedia information retrieval community. Using the experience obtained during this tool development, we designed and developed the visualization module for our GF-based frames classifier and polyp detector.

3.5.2 Visualization Module for Polyp Detection and Spotting

The visualization module for real-time polyp detection and spotting is designed to be integrated into the complete live EIR system pipeline. The primary aim of the EIR system is to provide live feedback to doctors, i.e., a computer-aided diagnosis in real-time. Thus, while the endoscopist performs the colonoscopy, the system analyzes the video frames that are captured by the colonoscope. In this visualization module, we combine the visual information from the endoscope with our marks to provide helpful information for the operating doctor. For the detection, we alter the frame borders and show the name of the detected finding in the auxiliary



(a) (b)

Figure 3.20: Online global-feature-based visual similarity search tool usage examples. The image in the center is the query image. The first six results of four queries based on four three global and one local features are shown around the query image.

area of the endoscope device monitor. For the implemented lesion localization spotting, we draw a cross on top of the localized findings (polyps in this system version). Additionally, we plot in the lower part of module's UI display additional information about the lesion detection performance including the polyp localization ground truth, per-frame polyp detection indicator and, most important for the visual detection performance verification, event recorder that depicts detection events, e.g., true positive (TP), false positive (FP), false negative (FN) and true negative (TN), for each and every processed frame. The visualization module together with the underlying detection and localization (spotting for polyps) subsystems is able to process a Full HD video stream with 30 FPS that meets our in real-time goal. An example of the graphical output of the live system is depicted in figure 3.21. The visualization module is implemented in C++ using the OpenCV library for video stream handling, and it is cross-platform supporting the Windows and Linux operating systems.

For the deep-learning-based detector, we implemented an additional visualization module especially designed to provide efficient integration with the Python-based DL subsystems. The designed Python wrapper provides seamless video frame import in a separate worker thread, execution of various TensorFlow-based lesion detectors and drawing of the detection results together with the input video frames in unified UI (see Figure 3.22). In this module, we put most of the efforts into making our TensorFlow detection code work in parallel with the visual data input and output (I/O), to be able to utilize simultaneously CPU and GPU resources for data I/O and analysis, respectively.

3.5.3 Visualization Module for Lesions Detection and Localization

Our most recent visualization module for real-time polyp detection and localization is designed in tight collaboration with experienced endoscopists with the primary aim of enabling integration with real endoscopic equipment installed in the hospitals' examination rooms. Despite

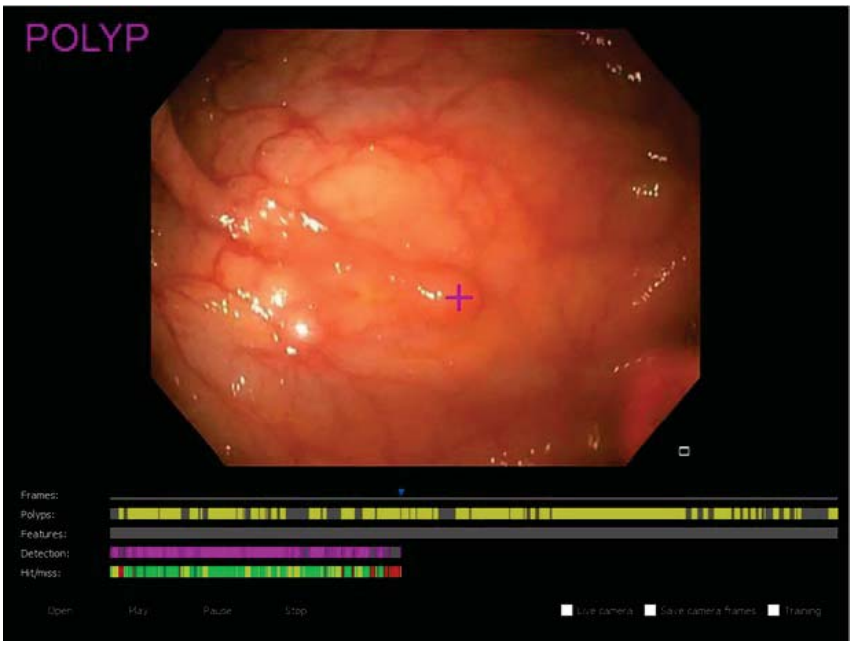


Figure 3.21: The visualization module for real-time polyp detection and spotting build upon our global-feature-based detection and hand-crafted local-feature-based polyp position finder approaches. It is able to process both recorded and live Full HD video stream from traditional colonoscope, highlight frames containing polyps and mark the recognized polyp location with a cross mark. The pink surrounding frame shows a positive detection. Plot in the lower part of UI shows the per-frame polyp presence ground truth, polyp detection indicator and TP/FP/FN/TN events recorder.

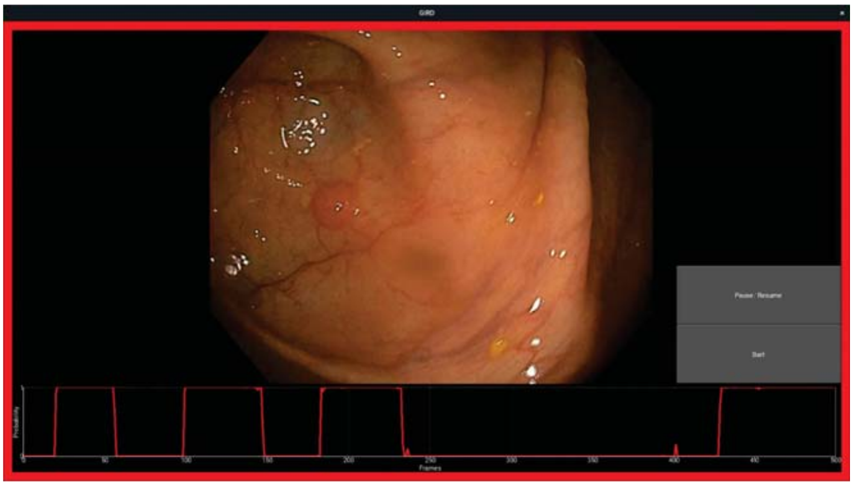


Figure 3.22: The visualization module for our deep-feature-based real-time polyp detection approaches. It is able to process Full HD live-captured video stream from traditional colonoscope and highlight frames containing detected lesions. The plot in the lower part of UI show the per-video-frame lesion detection probability.

visual feedback simplicity (see figure 3.23), its architecture supports input from Full HD live video endoscope sources and provides as low latency as possible in order to minimize the overall pipeline execution time for individual video frames. This is especially important for live examinations when endoscope and instrument movements are precisely controlled by only visual feedback on the primary operational display. During the initial clinical trials, we will display the visual detection and localization output on the auxiliary screen to avoid possibility of the video footage interruptions, thus the initial latency requirements are not as strict as they will be for the main trials with only one primary display with the integrated lesion detection and localization marking. Thus, for this EIR system version, we do not define the target processing latency, rather we set the minimum frame processing rate of the 15 FPS as enough for the initial live detection and localization system implementation. Nevertheless, this relatively low target processing speed is enough for live system evaluation in real-world conditions. On the other hand, it is not reducing system benefits for off-line endoscopy data processing, due to its independent processing of frames. For post-procedure or VCE data processing, the analysis is easily parallelized, resulting in a high EIR system scalability [96, 101].

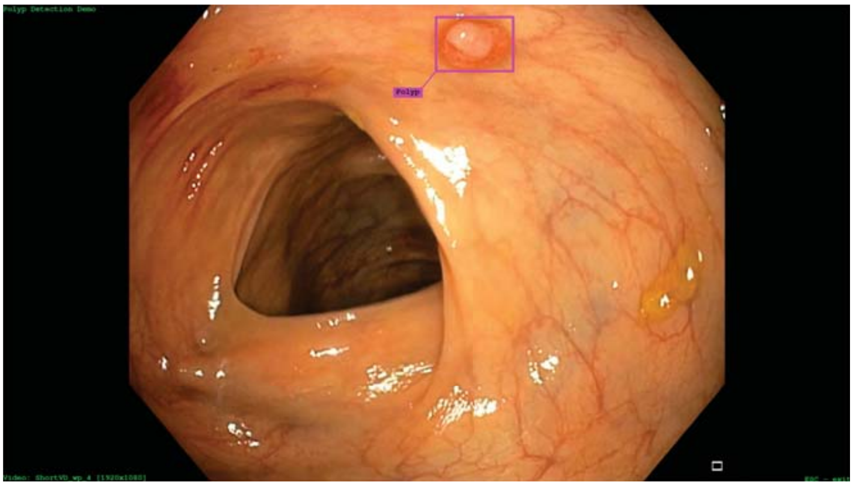


Figure 3.23: Near-to-real-time polyp detection and localization demo build upon our GAN-based detection and localization approach. The software processes recorded Full HD video stream from traditional colonoscope and highlights the exact polyp location in the particular frame. The marking is implemented as as a bounding box rectangle drawing over the source video frame. The achieved processing speed is in between 5 and 10 FPS depending on the used GPU acceleration hardware.

3.6 System Evaluation

In this section, we present the experiments that we conducted on the DeepEIR system. We tested the whole system and its individual subsystems in terms of usability, accuracy and data processing performance. The requirements of the system that we are evaluating are: (i) ability to handle big amounts of data during data collection and annotation phases; (ii) reaching real-time performance (being able to process 25-30 frames per second); (iii) achieving high detection and localization accuracy (at least equal to the best related approaches in table 2.2); and ability to visualize detection and localization results in a convenient way. All the experiments except for the shared GPU and extreme multi-core CPU-efficiency testing were conducted using consumer-grade computation equipment and general-purpose GPUs without utilization of specialized CNN-oriented accelerators.

3.6.1 Annotation Subsystem

We evaluated performance and usability of the annotation system during the exploration and annotation of our two datasets Kvasir and Nerthus. In the initial stages of the project, we mostly were processing and sorting the raw anonymized data received from hospitals' information systems manually. Despite the fact that it is not possible to fully avoid any manual annotation work during the dataset preparation and verification, the amount of work was tremendous, and it took several weeks to prepare the very first pre-version of the Kvasir dataset.

As the initial annotation-automation approach, we implemented a visual-feature-based sorting algorithm. First, we used our OpenSea tool to extract global image features from all the unsorted images. Next, we used the K-Means clustering algorithm from WEKA to build a set of clusters containing visually similar images in the different clusters. Finally, generated clusters were processed manually in order to select a small set of relevant images for the classes of diseases. This intermediate solution was, next, evolved into the hyperbolic-tree-based visualization and clustering tool. This tool was used for the further raw dataset exploration. The hyper-tree-based representation significantly improved our ability to explore the data collection, however, the graphical view's drawing performance was not sufficient to process the larger collections containing thousands of images. Thus, we continued to evolve the tool.

The resulting ClusterTag cluster-based visualization and annotation tool was especially designed with the big data collections in mind. The annotation automation was improved by introducing classification-based clustering capability. The user can easily improve the quality of clusterization by using a set of pre-selected images for each defined image class. The pre-selected (seed) images are then used as a training set for our classification methodology introduced in our OpenSea classifier. After model training, the remaining raw images are classified by OpenSea, and the classification results then used to make new clusters of pre-annotated images. This resulted in better cluster density, significantly reducing the amount of manual work required for dataset annotation.

To solve the issue of drawing performance, we used a set of techniques to support low-latency visual representation and give the best possible user-friendly experience to the annotators. The ClusterTag tool itself, as well as all the used libraries, is written in Java and, thus, it is a cross-platform solution that can be easily deployed on Windows, Linux and macOS. The drawing constrains introduced because of Java's cross-platform nature were resolved using the platform-targeted Lightweight Java Game Library, which is using OpenGL for hardware-accelerated painting. The access performance of the storage used for drawing data was improved by developing a high-speed custom image caching technique and background database update strategy. All together, our efforts to implement real-time drawing of big image collections resulted in an efficient visual core implementation. The screen update and redraw latency of 100 millisecond and less was measured for the big collection of 200.000 images of different resolution varying from QCIF up to Full HD. Further improvement of drawing performance can be achieved by porting the drawing core to C++ and implementing sub-scale images caching in GPU memory.

Participant	True Positive	False Positive	False Negative	Precision	Recall	F1 score
UNS-UCLAN	48	481	148	9.07	24.49	18.28
CuMedVis	31	167	165	15.75	15.81	15.77
CVC	33	163	163	16.84	16.84	16.84
Our EIR System	46	723	150	5.98	23.47	14.81
RUS	65	1558	131	4.00	33.16	13.50
SNU	8	188	188	4.08	4.08	4.08

Table 3.1: Results of the MICCAI 2015 polyp localization challenge [25].

3.6.2 Detection and Localization Subsystems

3.6.2.1 Evaluation Metrics

For the performance evaluation experiments, we used the following metrics precision (PREC), recall/sensitivity (SENS), specificity (SPEC), accuracy (ACC), F1 score (F1) and Matthew correlation coefficient (MCC). A detailed description and reasoning for the used metrics is given in paper XII. The detection performance metrics are computed frame-wise. The localization performance metrics are computed pixel- and block-wise depending on the approach being evaluated using the provided binary masks of the ground truth.

The data processing speed is measured in number of frames per second (FPS). For all the approaches we use the margin of 25 FPS as a border-line for the algorithm to be considered real-time-capable.

3.6.2.2 Polyps

The very first evaluation of our polyp detection and localization approach was performed by participating in the MICCAI 2015 Grand Challenge [25]. It this challenge, three different databases were used. Two publicly available databases were proposed for still-frame analysis, CVC-CLINIC and ETIS-LARIB. CVC-CLINIC [24] contains 612 SD frames and comprises 31 different polyps from 31 sequences. ETIS-LARIB [4] contains 196 HD frames and comprises 44 different polyps from 34 sequences. All the images contain at least one polyp. The ground truth consists of the polyp masks annotated by qualified endoscopists from the corresponding clinical institution. The last one is the closed and copyrighted ASU-Mayo Clinic Colonoscopy Video Database [1], which comprises a set of short and long colonoscopy videos, col- lected at the Department of Gastroenterology at Mayo Clinic, Arizona. This database consists of 38 different, fully annotated videos including frames with and without polyps.

The challenge consisted of two sub-tasks: polyp localization and polyp low-latency detection. The polyp localization sub-task is designed to find out if the proposed method can cope with variability of polyp appearance within a captured video-frame and, therefore, accurately determine the location of a polyp in the frame. The low-latency detection checks it the proposed method can detect a polyp in the frame and determine the delay from the first appearance of the polyp to the moment when it is detected.

Table 3.1 depicts the result for the polyp localization part based on the CVC-ClinicDB dataset. EIR was on the fourth place out of six. Based on the fact that our system is not built for only polyp detection, the achieved results were promising. It is also important to point out that the first three participants were organizers of the challenge and involved in the dataset collection. Table 3.2 gives an overview of the results for the detection latency part.

Participant	Latency in ms	F1
CuMedVis	6.66	26.40
Our EIR System	21	13.27
SNU	43.33	6.13
CVC	44.60	22.78
Rustad	235	11.47
ASU	417.5	20.84
UNS-UCLAN	0	0

Table 3.2: Results of the MICCAI polyp detection challenge. The table shows the detection latency in milliseconds and F1 score [25].

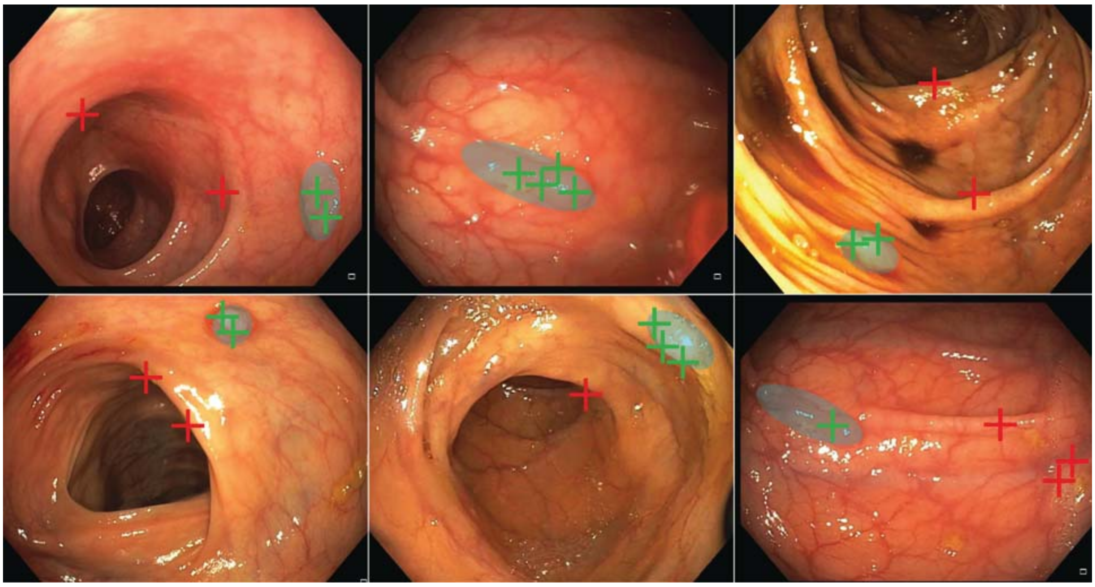


Figure 3.24: Polyp localization results generated by our first polyp localization and detection approach on the MICCAI 2015 dataset [25]. Light green ellipses depicts the polyp localization ground truth masks. Green and red crosses show the true positive and false positive polyp localization results, respectively. The localization algorithm was tuned to output exact four possible polyp locations per frame.

For the latency, EIR performed second best out of all participants. This is a very good result, and a positive confirmation of the real-time performance capability of EIR. It should also be mentioned that the approach of UNS-UCLAN is not able to distinguish between a frame with or without polyp.

Overall, the results of the challenge were positive for a system that is designed to be expandable with different diseases and use cases. We proved that we were able to compete and outperform other state-of-the-art approaches, which are designed for the specific problems of the challenge, without applying any adaptations or modifications to EIR or tuning our detection for the given dataset [25]. It is also important to point out that we participated in the MICCAI 2015 challenge at the early stage of EIR system development in order to validate our approaches under real-world conditions.

In the following stages of our work, polyp detection and localization were the main focus of this research, and the performance of polyp detection and localization has been gradually increased. The recent and most promising results were reached via our latest approach which uses a combined algorithm to address both polyp detection and localization at the same time. However, to be properly trained, the method requires the detailed ground truth masks for each and every training image used. Thus, to assess the method's performance, we used another publicly available dataset apart of our Kvasir and Nerthus datasets. This additional dataset is a part of MICCAI 2017 Grand Challenge [23] and it is publicly available for research purposes.

All-in-all, for the performance evaluation experiments, we use combinations of six different datasets, namely CVC-356 [23], CVC-612 [24], CVC-968, CVC-12k [23], Kvasir [95] and parts of Nerthus as the source of normal mucosa frames [94] (see Table 3.3 and Paper XV for the detailed datasets overview). The CVC-356 and CVC-612 datasets consist of 356 and 612 video frames, respectively. CVC-968 is a direct combination of CVC-356 and CVC-612. In these datasets, each frame that contains a polyp comes along with pixel-wise annotations. All three small CVC datasets are used for both training and testing the localization performance-evaluation experiments, and for the training only in the detection experiments. For all frame-wise polyp detection approaches, except for the GAN-based approach, we also added the $1,350$ frames of normal mucosa from the Nerthus dataset since normal mucosa examples for the negative class are required for our GF- and DF-based detection algorithms. The big CVC-12k dataset contains $11,954$ frames extracted from different videos, $10,025$ of them contain a polyp and $1,929$ show only normal mucosa. The polyps are not precisely annotated pixel-wise, but with an oval shape covering the approximated polyp body region (approximated annotation). For the Kvasir dataset, we included all the classes except for the dyed classes (in a real world scenario something dyed is already easily detected by the doctor) leading to a frame-wise annotated dataset containing $1,000$ frames with polyps, and $5,000$ without. The CVC-12k dataset is used as the test set for block- and frame-wise detection and the Kvasir dataset - for frame-wise detection approach evaluation.

All the images and video frames used in polyp localization and detection evaluation experiments are captured from standard endoscopic equipment and can contain some additional information fields related to the endoscopic procedure. Some types of the fields (see Paper XV for the details) integrated into resulting endoscopic frames can confuse detection and localization approaches, and lead to frame misclassification or false positive detection (captured frame with a polyp). To avoid these problems, we have implemented a simple frame preparation procedure that consists of three independent steps: a black border removal (including patient-related text fields), a green navigation localizer map masking and a captured still frame masking. All the removed and masked regions were excluded from further frame analysis. Moreover, due to a limited number of available frames with detailed ground truth masks, we implemented a data augmentation scheme used in the training procedure for the GAN-based approaches. For the presented evaluation, we used only rotation and flipping of frames. Rotation was performed independently with $20°$ steps. Together with the in-horizontal-direction-flipped frames, we added 35 new frames complementary to each original one.

Table 3.4 depicts the performance evaluation results for the GAN-based pixel-wise polyp segmentation approach. The best performance is achieved using the CVC-612 dataset for training, which means, more training data improves the final results. An interesting observation is that the precision is higher with CVC-356 as training data. This might be an indicator that more training data makes the model more general, but less accurate. All in all, the validation using these datasets indicates that the approach works well, and the proposed localization algorithm can perform efficiently even with a low number of available training samples. This is important for our medical use-case scenario with a high diversity of objects and a limited amount of

Dataset	Training	Test	# Frames	# Polyp frames	# Normal frames
CVC-356	X	X	1,706	356	1,350
CVC-612	X	X	1.962	612	1,350
CVC-968	X	X	2.318	968	1,350
CVC-12k	-	X	11,954	10,025	1,929
Kvasir	-	X	6,000	1,000	5,000
Nerthus	X	-	1,350	-	1,350

Table 3.3: Overview of the datasets used in the experiments. Kvasir and Nerthus are our own public datasets. CVC-968 is a combined dataset consist of CVC-356 and CVC-612 sets.

Test set	Run	Train set	PREC	SENS	SPEC	ACC	F1	MCC
CVC-612	LOC-356	CVC-356	0.819	0.619	0.984	0.946	0.706	0.684
CVC-356	LOC-612	CVC-612	0.723	0.735	0.981	0.965	0.729	0.710

Table 3.4: Validation results of the in-frame pixel-wise polyp areas segmentation (localization) approach evaluated using different combinations of the CVC-356 and CVC-612 sets for training and testing.

Run	PREC	SENS	SPEC	ACC	F1	MCC
LOC-Xception	0.584	0.257	0.972	0.880	0.357	0.333
LOC-VGG19	0.232	0.406	0.800	0.750	0.295	0.166
LOC-ResNet50	0.536	0.248	0.968	0.875	0.340	0.306

Table 3.5: Performance of the block-wise polyp localization (LOC) via detection approaches reported per method and used training data. Training and testing are performed using the CVC-968 and CVC-12k datasets, respectively. See Paper XV for the detailed results.

annotated data available.

The results for the block-wise polyp location approaches are presented in Table 3.5. The performance results obtained are especially interesting since all the approaches presented are trained with small amounts of training data without any negative examples (no normal mucosa frames at all). Furthermore, the CVC-12K dataset is heavily imbalanced, which makes it harder to achieve good results. For block-wise location via detection, the LOC-Xcept approach performs best for all the different training set sizes. It also indicates that a larger training dataset can lead to better results. The results for the LOC-ResNe approach confirm this with significant improvements when the training dataset size is increased. This is something that should be investigated in the future. Additionally, the algorithm used to combine the results on different sub-frames into one can be improved by, for example, using another machine learning algorithm to learn the best combinations.

The frame-wise polyp detection results can be found in Table 3.6. All approaches are trained on CVC-356, CVC-612 and CVC-968 training datasets and tested on the CVC-12k and Kvasir datasets. All in all, the GAN approach performs best on both datasets and within all variations of training datasets. The performance on the Kvasir dataset is better than on the CVC-12k dataset which is surprising since the Kvasir data is completely different from the CVC training data. Moreover, frames in the Kvasir dataset are captured using different and various hardware. This is a strong indicator that the approach is able to create a general model that is not just working

Test set	Run	PREC	SENS	SPEC	ACC	F1	MCC
Kvasir	GAND-Kvasir	0.736	0.746	0.946	0.913	0.741	0.689
	GFD-Kvasir	0.225	0.859	0.409	0.484	0.357	0.208
	RTD-Xception-Kvasir	0.459	0.256	0.939	0.825	0.328	0.251
	RTD-VGG19-Kvasir	0.231	0.320	0.842	0.774	0.268	0.142
	RTD-ResNet50-Kvasir	0.248	0.877	0.469	0.537	0.387	0.262
	YOLOD-Kvasir	0.530	0.559	0.901	0.844	0.544	0.450
CVC-12k	GAND-CVC-12k	0.906	0.912	0.510	0.847	0.909	0.428
	GFD-CVC-12k	0.835	0.854	0.125	0.737	0.845	-0.020
	RTD-Xception-CVC-12k	0.899	0.690	0.600	0.676	0.781	0.224
	RTD-VGG19-CVC-12k	0.232	0.406	0.800	0.750	0.295	0.166
	RTD-ResNet50-CVC-12k	0.870	0.303	0.766	0.378	0.450	0.057
	YOLOD-CVC-12k	0.932	0.641	0.757	0.660	0.759	0.296

Table 3.6: Results for the frame-wise polyp detection approaches, namely multi-class global-feature-based (GFD), deep-learning-based with random tree (RTD) final classifier, GAN-based (GAND) and YOLOv2-based (YOLOD). We used the CVC-12k and Kvasir dataset as independent test sets. Training of all the approaches is performed using the combined CVC-968 dataset consist of CVC-356 and CVC-612 sets. See Paper XV for the detailed results.

well on the given data and that the CVC-12k dataset is very challenging. Some of the difficulties we could observe are for example screens in screens that show different parts of the colon, out of focus, frame blur, contamination, etc. (see for example Figures 3.18 and 3.26). From the RTD approaches, Xception-based has the best overall performance, and it performs best on the CVC-12k dataset. The ResNet50-based method reaches best performance for the Kvasir dataset, but is still far away from the GAN approach (MCC 0.262 versus 0.689). The GFD approach did not perform well on the CVC-12k dataset and could not make sense of the data. This is indicated by only negative MCC values which basically means no agreement. On the Kvasir dataset, it performed much better and could even outperform RTD VGG19-based approach. Overall, the RTD approaches with VGG19 performed worse than all other approaches. The reason could be that the general hyper-parameters that we collected using optimization did not work well for the VGG19 architecture.

In order to compare our detection approaches to the state-of-the-art, we also evaluated one of the recent and promising object detection CNNs called YOLOv2 [107]. The YOLOv2 model is able to detect objects within a frame and to provide an object's localization box and a probability value for the object detection. We trained YOLOv2 with the CVC-968 dataset using an appropriate conversion from ground truth masks to surrounding object boxes, as required by YOLOv2. The training was performed from scratch with the default model parameters. The trained YOLOv2 model showed relatively high performance with an MCC value of 0.450 and 0.296 for the Kvasir and CVC-12k sets, respectively, and was able to outperform all tested approaches except for the GAN-based solution. Nevertheless, the performance of the well-developed and already fine-tuned YOLOv2 model is significantly lower than our new GAN-based detection-via-localization approach.

Table 3.7 depicts the performance evaluation results for the GAN-based pixel-wise localization (segmentation) approach using two polyp datasets with detailed ground truth masks available: CVC-356 and CVC-612. In this experiment, we performed a cross-validation using these two datasets. The best performance is achieved (as it was expected), using the bigger CVC-612 dataset for training. Here, we achieved a well-balanced localization performance with the high overall measures F1 of 0.729 and MCC of 0.710. An interesting discovery of this experiment is that our localization algorithm can still perform very efficiently (F1 of 0.706 and MCC of 0.684) even when trained using the small amount of training data (CVC-365 contains only 356 images of polyps). This is a vital property for our medical use-case scenario with a high diversity of objects and a limited amount of annotated data available. Figure 3.25 shows the representative example of the polyp localizer output. The pixel-wise probability mask shows the possible localization of the polyp body's pixels and it conforms well with the ground truth. Comparing to our initial polyp localization, the GAN-based approach can easily distinguish between normal intestinal folds and polyp-affected tissue by learning the tiny local image features and shape properties.

Another experiment shows our approach to the common case of coarse ground truth available for the data. Here we use our block-wise location via detection approach. The performance results presented in table 3.8. The best performance with F1 score of 0.357 and MCC of 0.333 was achieved using the CVC-968 dataset. The interesting insight is that the algorithm was

Test set	Run	Train set	PREC	SENS	SPEC	ACC	F1	MCC
CVC-612	LOC-356	CVC-356	0.819	0.619	0.984	0.946	0.706	0.684
CVC-356	LOC-612	CVC-612	0.723	0.735	0.981	0.965	0.729	0.710

Table 3.7: This table depicts performance of the in-frame pixel-wise polyp localization (segmentation) approach evaluated using different combinations of the CVC-356 and CVC-612 datasets for training and testing.

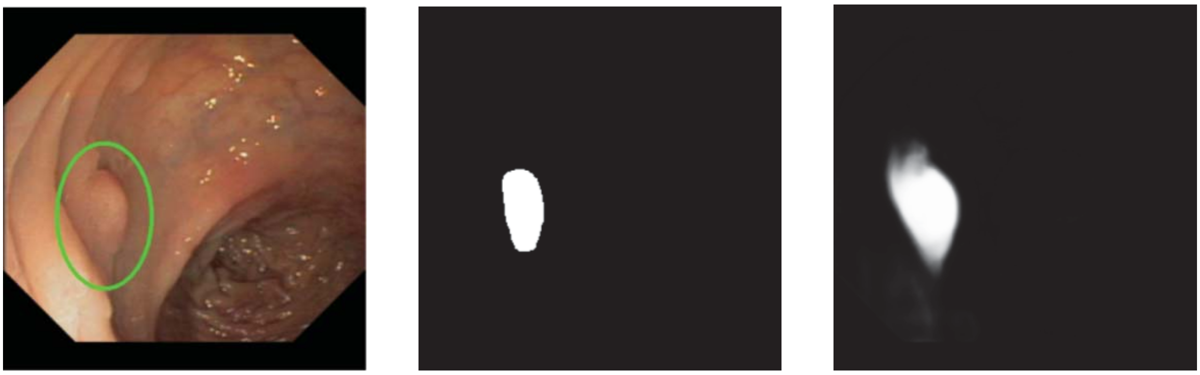

(a) Input image
(b) Ground truth mask
(c) Polyp localization probability mask

Figure 3.25: The example ot the polyp localization mask generated by our GAN-based polyp localization approach. The base polyp localizer generates the pixels-wise probability mask shows the possible localization of the polyp body's pixels. The green ellipse highlights the polyp body for illustration purposes only. The resulting localization mask conforms good with the ground truth.

Training set	PREC	SENS	SPEC	ACC	F1	MCC
CVC-356	0.475	0.203	0.966	0.868	0.285	0.250
CVC-612	0.528	0.289	0.961	0.874	0.374	0.328
CVC-968	0.584	0.257	0.972	0.880	0.357	0.333

Table 3.8: This table depicts performance of the block-wise localization via detection approach for the CVC-12K dataset reported for different training data used.

Test set	Training set	PREC	SENS	SPEC	ACC	F1	MCC
Kvasir	CVC-356	0.715	0.751	0.940	0.909	0.732	0.677
	CVC-612	0.595	0.803	0.891	0.876	0.684	0.619
	CVC-968	0.736	0.746	0.946	0.913	0.741	0.689
CVC 12k	CVC-356	0.967	0.624	0.888	0.667	0.758	0.378
	CVC-612	0.934	0.609	0.778	0.636	0.737	0.286
	CVC-968	0.906	0.912	0.510	0.847	0.909	0.428

Table 3.9: This table depicts performance of the frame-wise polyp detection approach. We used different small training sets and the CVC-12k and Kvasir dataset as independent test sets.

trained with a small amount of training data without any negative samples (no normal mucosa frames is presented). Furthermore, the CVC-12K dataset is heavily imbalanced which also makes it harder to achieve good results.

The frame-wise detection results can be observed in Table 3.9. All approaches are trained on CVC-356, CVC-612 and CVC-968 training datasets and tested on the CVC-12k and Kvasir datasets. We reached an F1 score of 0.741 and an MCC score of 0.689 for the Kvasir test dataset. For the CVC-12k test set, we reached an F1 score of 0.909 and an MCC score of 0.428.

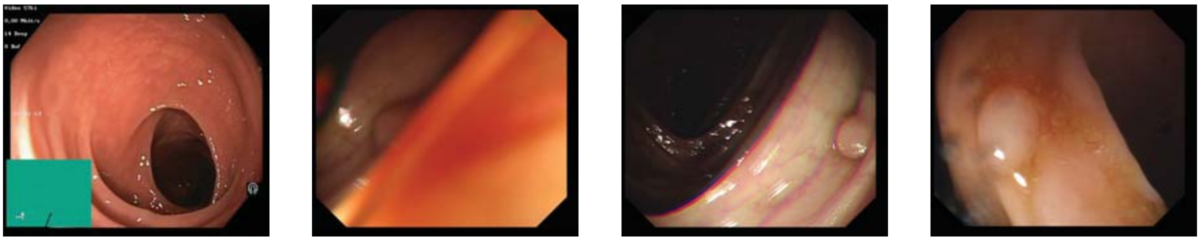
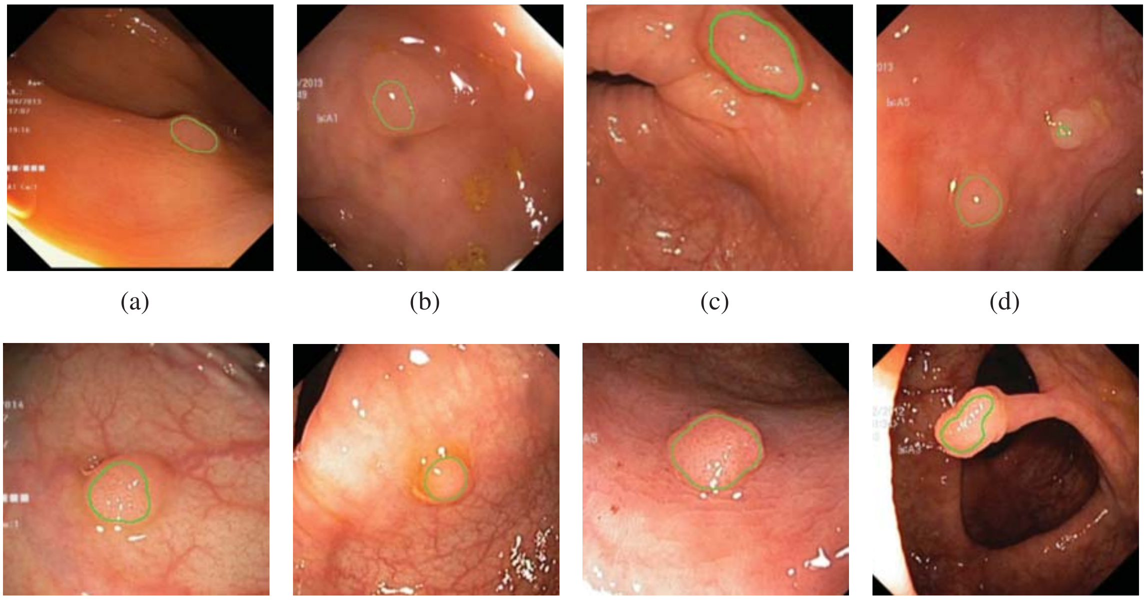

(a) Overlay image
(b) Blurry frame
(c) Colors shift
(d) Lens contamination

Figure 3.26: Example of difficult images in the test dataset: a significant frame blur caused by camera motion (a), a color components shift caused by the temporary signal failure (b) and an out-of-focus frame contains also contamination on the camera lens (c). Images taken from the CVC-12k [23].

3.6.2.3 Angiectasia

After a successful evaluation of the GAN-based polyp detection and localization approach, we decided to check whether is it flexible enough and how it can be extended to other GI tract lesions. To test as meaning as possible, we chose the angiectasia lesion in a combination with the VCE-based diagnostic method. In contrast to polyps, angiectasia is a flat mucosa lesion. The main feature differentiating it from the surrounding normal tissue is color. However, the



(a) (b) (c) (d)

(e) (f) (g) (h)

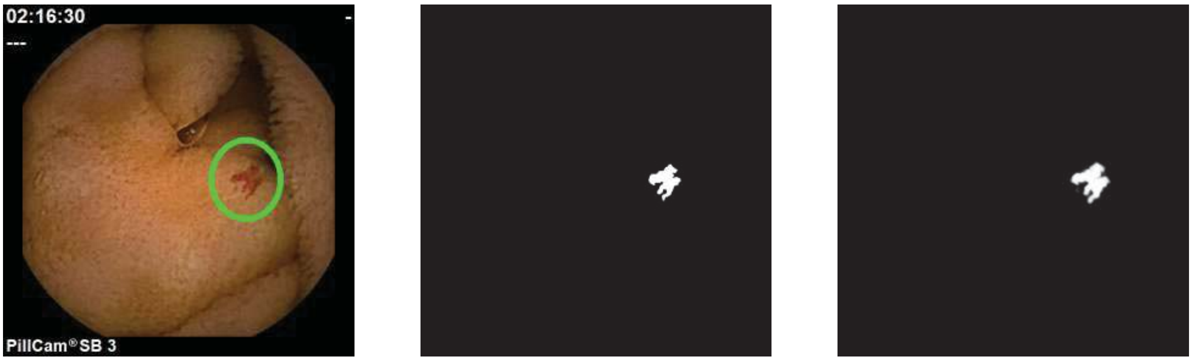
Figure 3.27: Examples of the detection and in-frame localization of the different polyps in the video frames captured by various vendors' traditional colonoscopy equipment. Green contour depicts the detected polyp and the localized main polyp body area.

size of angiectasia-affected mucosa areas can be rather small and they still need to be detected and localized.

The data used for all the angiectasia detection and localization experiments is from the GIANA 2017 challenge [22], and it is publicly available for research purposes. The data consists of training (development) and test frame sets. The training set consists of 600 fully annotated frames from VCEs (300 with angiectasia and 300 without). The frames with angiectasia also have a pixel-wise ground truth (GT) mask depicting the exact lesion location inside each frame that allows both pixel-wise localization and frame-wise detection experiments. The test set consists of 600 unannotated frames. In order to perform validation and performance evaluation of the developed detection algorithm, we annotated the test set frame-wise with the help of an experienced researcher with a background in medical pathology diagnosis. The 600 frames from the development set are used for training and the 600 frames (300 with angiectasia and 300 normal) from the test set for verification. The advantages of the used dataset are (i) the number of images (compared to related work, this is the largest one for VCEs), (ii) the even split between positive and negative examples and (iii) that it is publicly available making it easy to compare different approaches.

Table 3.10 shows the results for the GAN localization algorithm (see figure 3.28(b) and 3.28(c) for a comparison between the GT and the output of the GAN). The localization metrics are calculated pixel-wise using the provided GT masks. On average, sensitivity and specificity are above the 85% margin recommended for a real clinical settings. This can be seen as very good results since we perform pixel-wise evaluation. The processing speed for the GAN approach is 1.5 FPS.

The frame-wise detection performance of the GAN approach for the development set is



(a) Input frame (b) Ground truth mask (c) Segmentation mask

Figure 3.28: Example of an angiectasia lesion marked with a green circle (a), a corresponding ground truth mask (b) and a segmentation mask generated using our GAN-based approach (c). Image taken from the GIANA dataset [22].

PREC	SENS	SPEC	ACC	F1	MCC
0.859	0.880	0.999	0.999	0.869	0.869
±0.020	±0.018	±0.001	±0.001	±0.015	±0.015

Table 3.10: This table depicts ten-fold cross-validation results of the pixel-wise GAN-based angiectasia localization approach (the 95% confidence intervals are reported). See Paper XIV for the detailed results.

PREC	SENS	SPEC	ACC	F1	MCC
1.000	0.987	1.000	0.993	0.993	0.987
±0	±0.011	±0	±0.005	±0.005	±0.011

Table 3.11: This table depicts ten-fold cross-validation results of the angiectasia frame-wise detection using the GAN approach (the 95% confidence intervals are reported). See Paper XIV for the detailed results.

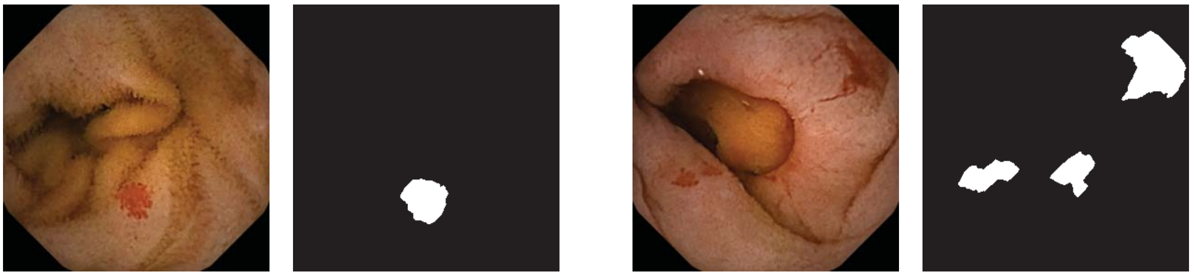
presented in Table 3.11. The detection outperforms significantly the 85% requirements. Both result sets are strong indicators that our GAN approach performs well for the tasks of angiectasia localization and detection.

Finally, in Table 3.12, we report the frame-wise detection performance on the test set for all our runs. All tested approaches outperform the ZeroR baseline, but most of them do not even come close to the 85% margin for clinical use. The handcrafted features outperform the VGG19 and InceptionV3 approaches but not the RestNet50. Of the classifiers, LMT performs best most of the time, followed by RF. The best performing not-GAN approach is *AUG DF ResNet50 FEA + LMT*. The GAN approach achieves superior performance compared to all other detection methods for the frame-wise detection with a sensitivity of 98% and a specificity of 100%.

The best processing speed is reached by the GF approach using RT. In terms of fastest speed and best classification performance, *AUG DF ResNet50 CON + RF* performs best with a sensitivity of 78.7% , a specificity of 78.7% and a processing speed of 78 FPS. The processing speed of the GAN method for detection is the lowest with 1.5 FPS.

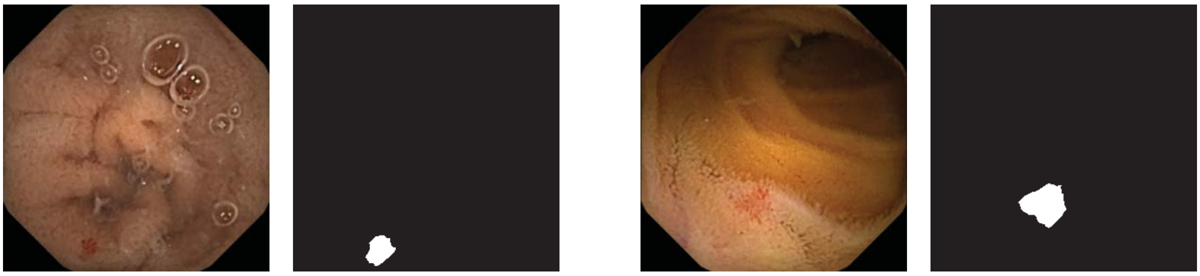
Approach	PREC	SENS	SPEC	ACC	F1	MCC	FPS
GF+LMT	0.695	0.680	0.680	0.680	0.674	0.375	80
DF ResNet50 CON+LMT	0.734	0.732	0.732	0.732	0.731	0.465	53
DF ResNet50 FEA+LMT	0.748	0.738	0.738	0.738	0.736	0.486	46
DF VGG19 CON+LMT	0.545	0.545	0.545	0.545	0.544	0.090	32
DF VGG19 FEA+LMT	0.525	0.525	0.525	0.525	0.525	0.050	29
DF InceptionV3 CON+LMT	0.663	0.663	0.663	0.663	0.663	0.327	37
DF InceptionV3 FEA+LMT	0.533	0.533	0.533	0.533	0.533	0.067	30
AUG GF+LMT	0.627	0.625	0.625	0.625	0.624	0.252	80
AUG DF ResNet50 CON+LMT	0.765	0.763	0.763	0.763	0.763	0.529	53
AUG DF ResNet50 FEA+LMT	0.797	0.788	0.788	0.788	0.787	0.585	46
GAN	**1.000**	**0.980**	**1.000**	**0.990**	**0.990**	**0.980**	**1.5**
Baseline (ZeroR)	0.250	0.500	0.500	0.500	0.333	0.000	-

Table 3.12: Results for the angiectasia frame-wise detection approaches evaluated with the annotated test set. See Paper XIV for the detailed results.

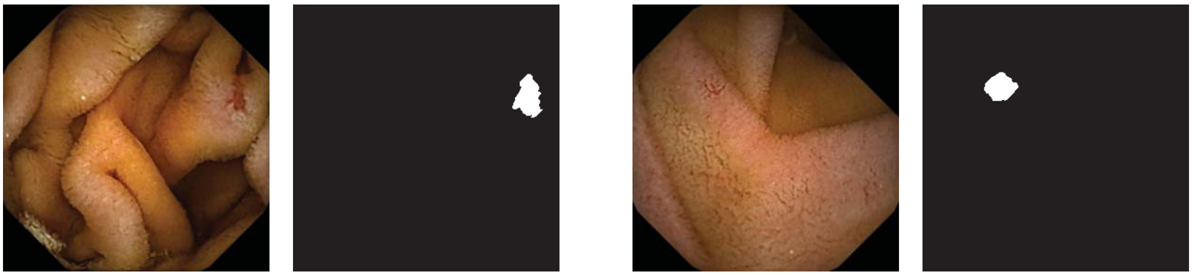


(a) Input VCE-frame (b) Localization result (c) Input VCE-frame (d) Localization result

Figure 3.29: Examples of the detection and in-frame localization of the clearly visible angiectasia areas.



(a) Input VCE-frame (b) Localization result (c) Input VCE-frame (d) Localization result



(e) Input VCE-frame (f) Localization result (g) Input VCE-frame (h) Localization result

Figure 3.30: Examples of the detection and in-frame localization of the partially obscured, tiny and hard-to-spot angiectasia areas.

		Detected class					
		Blurry	Cecum	Normal	Polyps	Tumor	Z-line
Actual class	Blurry	**250**	0	0	0	0	0
	Cecum	0	**183**	64	3	0	0
	Normal	0	34	**197**	19	0	0
	Polyps	1	17	45	**183**	4	0
	Tumor	0	0	1	4	**245**	0
	Z-line	0	0	0	0	0	**250**

Table 3.13: A confusion matrix for the six-classes detection performance evaluation for the Deep-EIR detection subsystem

3.6.2.4 Multi-Class Detection

Multi-class evaluation of our detection approach was performed using two different datasets. The first one is Kvasir, which we consider as a core detection performance evaluator. The second one is Medico, which introduces more classes of findings comparing to Kvasir and represents the real-world use-case scenario in terms of data amount and imbalance.

Kvasir dataset

We performed the core multi-class detection performance evaluation based on the first version of our public dataset Kvasir. From the whole dataset, we randomly selected 50 different frames of 6 different classes (see See Paper XI for the details): blurry frames, cecum, normal colon mucosa, polyps, tumor, and Z-line. The selected frames were used to create 10 separate datasets, each containing training and test subsets with equal numbers of images. Training and test subsets were created by equally splitting random-ordered frame sets for each of the 6 classes. The total number of frames used in this evaluation is 300: 150 in the training subsets and 150 in the test subsets. Each training and test subset contains 25 images per class. Multi-class classification is then performed on all 10 splits and then combined and averaged. Following this strategy, an accurate enough estimation about the performance can be made even with a smaller number of images.

First, we evaluated Deep-EIR that implements the deep learning neural network multi-class detection approach. Table 3.13 shows the resulting confusion matrix. The detailed performance metrics presented in table 3.14 and the results can be considered as good, they confirm that Deep-EIR performs well. All blurry and Z-line frames were classified correctly. Cecum and normal colon mucosa were often cross-mis-classified, which is a normal behavior, because from a medical point of view, normal colon mucosa is part of the cecum, and under real-world circumstances, this would not be a relevant mistake. Interesting polyps and tumors were detected correctly in most cases, as well as the Z-line landmark, which is important for our medical use case.

Second, we performed an evaluation of the multi-class global-feature-based EIR, which implements a global-feature multi-class detection approach. The multi-class global-feature-based EIR classifier allows us to use a number of different global image features for the classification. The more image features we use, the more precise the classification becomes. We generated indexes containing all possible image features for all frames of all different classes of findings from our training and test dataset. These indexes were used for multi-class classification, different performance measurements and also for leave-one-out cross-validation. Using our detection

	True Pos.	True Neg.	False Pos.	False Neg.	Recall (Sensitivity)	Precision	Specificity	Accuracy	F1 score
Blurry	250	1249	1	0	100.0%	99.6%	99.9%	99.9%	**99.8%**
Cecum	183	1199	51	67	73.2%	78.2%	95.9%	92.1%	**75.6%**
Normal	197	1140	110	53	78.8%	64.2%	91.2%	89.1%	**70.7%**
Polyps	183	1224	26	67	73.2%	87.6%	97.9%	93.8%	**79.7%**
Tumor	245	1246	4	5	98.0%	98.4%	99.7%	99.4%	**98.2%**
Z-line	250	1250	0	0	100.0%	100.0%	100.0%	100.0%	**100.0%**
Overall	*1308*	*7308*	*192*	*192*	*87.2%*	*87.2%*	*97.4%*	*95.7%*	***87.2%***

Table 3.14: Performance evaluation of the six-classes detection for the Deep-EIR detection subsystem

		Detected class					
		Blurry	Cecum	Normal	Polyps	Tumor	Z-line
Actual class	Blurry	**250**	0	0	0	0	0
	Cecum	0	**226**	21	3	0	0
	Normal	0	85	**165**	0	0	0
	Polyps	0	10	8	**226**	6	0
	Tumor	0	0	0	8	**242**	0
	Z-line	0	0	0	0	0	**250**

Table 3.15: A confusion matrix for the six-classes detection performance evaluation for the multi-class global-feature-based EIR detection subsystem

	True Pos.	True Neg.	False Pos.	False Neg.	Recall (Sensitivity)	Precision	Specificity	Accuracy	F1 score
Blurry	250	1250	0	0	100.0%	100.0%	100.0%	100.0%	**100.0%**
Cecum	226	1155	95	24	90.4%	70.4%	92.4%	92.1%	**79.2%**
Normal	165	1221	29	85	66.0%	85.1%	97.7%	92.4%	**74.3%**
Polyps	226	1239	11	24	90.4%	95.4%	99.1%	97.7%	**92.8%**
Tumor	242	1244	6	8	96.8%	97.6%	99.5%	99.1%	**97.2%**
Z-line	250	1250	0	0	100.0%	100.0%	100.0%	100.0%	**100.0%**
Overall	*1359*	*7359*	*141*	*141*	*90.6%*	*90.6%*	*98.1%*	*96.9%*	***90.6%***

Table 3.16: Performance evaluation of the six classes detection for the multi-class global-feature-based EIR detection subsystem

system, the built-in metric functionality can provide information on the different performance metrics for benchmarking. Further, it provides us with the late fusion of all the selected image features and performs the selection of the exact class for each frame in test dataset. Table 3.15 shows the resulting confusion matrix, which shows, like the Deep-EIR results, that the global feature-based detection approach performs well, too. Again, all blurry and Z-line frames were classified correctly. Cecum and normal colon mucosa were sometimes cross-misclassified. Polyps and tumors were detected correctly in most cases. The detailed performance metrics are presented in table 3.16 and can also be considered as good.

The comparison of these two approaches shows that both approaches have an equal excellent overall F1 score of 100% in Z-line detection. The global-feature approach with the 100% F1 score outperforms the neural network approach by a small margin in blurry frame detection.

The neural network F1 score detection for tumors is 98.2%, which is 1% better than the global-feature approach. Detection of other classes is better for the global-feature approach, giving the F1 scores of 79.2% and 74.3% for cecum and normal mucosa. Most importantly for our case study, polyp detection performed much better using the global-feature approach, giving the 92.8% F1 score (13.1% better than the neural network approach).

The performance evaluation of the cross-validation for both multi-class classification approaches (see table 3.17) confirms the high stability of the models used for the classification.

Our experimental comparison of the Deep-EIR and the global-feature-based EIR of the detection system shows clearly that the global-feature approach outperforms the deep learning neural network approach and gives better accuracy for almost all target detection classes (except several cases of misclassification of tumors) in conjunction with high 92.8% and 97.2% F1 scores for the most important findings: polyps and tumors. Moreover, when a sufficiently large training dataset covering all possible detectable lesions of the GI tract is used, the proposed global-feature approach for multi-class detection requires relatively little time for training [116] compared to days and weeks for the deep learning neural network approach. However, this conclusion is valid only for a well-balanced datasets which contain a fairly high amount of training data for each class and has clearly visually distinguishable classes, e.g. landmarks, fecal content, cancer, etc. Thus, our GF-based detection approach can be used as a fast-to-compute pre-classifier which allows the further selection of more precise, but slower classification algorithms.

Medico dataset

The dataset used for the further evaluation of multi-class detection algorithms consists of $14,033$ GI tract images with different resolutions (from 720x576 up to 1920x1072 pixels) that are annotated and verified by experienced medical doctors (endoscopists) for the ground truth. It includes 16 classes, showing anatomical landmarks, pathological and normal findings or endoscopic procedures in the GI tract, with different numbers of images for each class, split into development (training) and testing sets. The anatomical landmarks are *normal-z-line*, *normal-pylorus*, *normal-cecum*, *retroflex-rectum*, *retroflex-stomach*, while the pathological findings include *esophagitis*, *polyps* and *ulcerative-colitis*. The pre-, under- and post-surgery findings are the *dyed-lifted-polyps*, the *dyed-resection-margins* and the *instruments*. Additional classes include normal tissue with or without stool contamination, namely the *colon-clear*, the *stool-inclusions* and the *stool-plenty*, as well as some image classes that are not usable for diagnosis, namely the *blurry-nothing* and the *out-of-patient*.

For our experiments, we divided all the data onto development and test datasets consisting of $5,293$ images and $8,740$ images, respectively. We decided for an unequal split to reflect the

Approach	Mean absolute error	Root mean squared error	Relative absolute error, %	Root relative squared error, %
Deep-EIR	0.07284	0.20574	26.21936	55.21434
Multi-class global-feature-based EIR	0.09242	0.19644	33.2672	52.7148

Table 3.17: Performance evaluation of the cross-validation for the Deep-EIR and the multi-class global-feature-based EIR detection subsystems

Class	Training samples	Testing samples
blurry-nothing	176	37
colon-clear	267	1065
dyed-lifted-polyps	457	556
dyed-resection-margins	416	564
esophagitis	444	556
instruments	36	273
normal-cecum	416	584
normal-pylorus	439	561
normal-z-line	437	563
out-of-patient	4	5
polyps	613	374
retroflex-rectum	237	192
retroflex-stomach	398	397
stool-inclusions	130	506
stool-plenty	366	1965
ulcerative-colitis	457	524

Table 3.18: The per-class-contents of the training and test dataset used for the multi-class detection algorithms evaluation. This dataset was used for the Medico task at MediaEval 2018 contest [100].

real-world conditions in the medical use-case area where the amount of training data is typically less than the data forming the real examinations. Also both datasets are heavily unbalanced in terms of number of samples per class, which reflects the real practice in hospitals while doctors tend to collect only selected classes of images, where giving no attention to, for example, normal findings and routine objects like stool. Thus, the number of images per class in the sets can vary from a few to thousands of images (see Table 3.18 for the details).

The initial experimental studies showed that the our detection model is able to efficiently extract high-level features from the given medical images, and it converges quickly during the retraining process with sufficient classification performance. However, due to a heavily imbalanced training dataset and despite training data augmentation, the detection performance of some classes was not good enough. To solve this, we implemented an additional training dataset balancing procedure that performs equalization of the training set by extensive random augmentation of the training samples for the under-filled classes, like *instruments*, *blurry*, etc. This nearly doubled the number of training samples allowing for better classification performance for the classes with a low number of images provided. An additional classifier output post-processing step was implemented in order to address the different importance of the different classes as it was stated in the Medico task dataset description [100]. Specifically, we performed the prioritized selection of the resulting output class for each image based of the model's probability output. This was implemented as the selection of the first class with the detection probability higher than a set threshold from the array of classes sorted in order of their importance.

For the final evaluation of our detection approach on the Medico dataset, we used two separate models trained on the different datasets. The first model was trained on the training set created from the development set using the common rotation-scale-shift data augmentation procedure. The trained model was used to process the task's test set, and the classification output was post-processed using the prioritized classification selector with four different probability threshold settings from 0.75 to 0.1, resulting in the runs #2 - #5. For run #1, we used the

Run	TP	TN	FP	FN	REC	SPE	PRE	ACC	F1	MCC	RK
A1	474	8122	72	72	0.824	0.991	0.828	0.984	0.815	0.812	**0.854**
A2	474	8122	72	72	0.823	0.991	0.828	0.984	0.814	0.811	**0.854**
A3	470	8117	76	76	0.817	0.991	0.819	0.983	0.807	0.803	**0.845**
A4	440	8087	107	107	0.774	0.987	0.771	0.976	0.756	0.752	**0.786**
A5	333	7981	213	213	0.664	0.974	0.646	0.951	0.601	0.605	**0.582**
E1	469	8117	77	77	0.765	0.991	0.729	0.982	0.743	0.737	**0.844**
E2	469	8117	77	77	0.765	0.991	0.728	0.982	0.743	0.737	**0.844**
E3	465	8112	82	82	0.758	0.990	0.722	0.981	0.736	0.729	**0.835**
E4	430	8077	117	117	0.709	0.986	0.677	0.973	0.679	0.674	**0.766**
E5	313	7960	233	233	0.546	0.971	0.607	0.947	0.504	0.510	**0.544**
ZR	34	7681	512	512	0.063	0.938	0.004	0.883	0.007	0.0	0.0
RD	35	7682	511	511	0.057	0.938	0.064	0.883	0.055	0.001	0.002
TR	546	8193	0	0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 3.19: Classification performance evaluation for the detection models, trained using the augmented (A) and size-equalized (E) training sets including ZeroR (ZR), Random (RD) and True (TR) baseline classifiers. Runs #1 corresponds to the non-prioritized classification, while runs #2 - #5 corresponds to the 0.75 to 0.1 classification probability threshold level.

max-probability selector without class prioritization. The results using the first model were considered as speed runs. The second model was trained using the equalized training set, and the same rules for the five run generation were considered as the detection run.

The computed performance numbers are depicted in table 3.19. All the runs significantly outperform the ZeroR and Random baselines and show good classification performance. All the runs that utilize the equalized training set have slightly better classification performance. Surprisingly, the introduced prioritized classification method did not result in improved detection performance, neither for the original nor for the equalized training sets. With the threshold of 0.75, the classification performance is equal to the non-prioritized runs. It means that the trained classifier is performing as well as it can, and additional re-classification using the class priorities does not make sense for this particular dataset. However, it still can be potentially interesting for bigger datasets or a higher number of classes. The best performing run was the detection run #1 generated using the equalized training set and non-prioritized classifier with the classification performance of 0.854 for Rk statistic (MCC for k different classes). The confusion matrix for this run is depicted in table 3.20, and the class imbalance and corresponding training and classification challenges can be easily observed. The most challenging class was *Instruments*. That is mostly caused by the different shapes, positions and visibilities of the instruments in the images. There was also a number of misclassification cases for the *Dyed* classes as well as for *Esophagitis* and *Normal Z-line* classes.

3.6.3 Detection Subsystem Processing Speed Optimization

Despite the demonstrated high lesion detection performance, the overall data processing speed of the complete EIR system pipeline was not enough for both implementation of simultaneous detection and localization of multiple diseases, and not for implementation of population-wide mass-screening of GI tract diseases, either. In our research, we target a general well-scalable system for automatic analysis of GI tract videos with high detection accuracy, abnormality localization in the video frames and better than real-time performance, thus it is important to

		Detected class															
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Actual class	A	**459**	2	1	1	5	0	1	0	54	0	13	13	1	7	0	7
	B	2	**388**	77	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	0	145	**451**	0	0	0	4	0	0	0	1	0	0	0	0	0
	D	0	0	0	**406**	81	0	0	0	1	0	4	0	0	0	0	26
	E	0	0	0	115	**462**	0	0	0	0	0	0	1	1	1	0	17
	F	0	0	0	0	1	**2**	0	0	0	0	0	0	0	0	0	0
	G	3	18	27	0	0	0	**548**	0	0	0	2	0	2	1	4	1
	H	10	1	0	5	2	0	0	**498**	98	0	3	1	24	0	0	6
	I	14	0	0	5	1	0	0	0	**1771**	0	5	2	1	3	0	7
	J	2	0	0	0	0	3	0	1	7	**37**	0	0	2	1	0	0
	K	22	1	6	17	2	0	7	1	8	0	**316**	14	1	9	0	64
	L	19	0	0	2	6	0	1	0	16	0	22	**551**	8	3	0	4
	M	3	0	1	1	0	0	0	6	4	0	5	1	**1025**	1	0	6
	N	8	0	0	3	4	0	0	0	3	0	2	1	0	**160**	4	8
	O	0	1	0	0	0	0	0	0	2	0	0	0	0	5	**387**	1
	P	0	0	0	1	0	0	0	0	1	0	1	0	0	1	2	**126**

Table 3.20: Confusion matrix for the run A1 depicted in table 3.19. The classes are Ulcerative Colitis (A), Esophagitis (B), Normal Z-line (C), Dyed and Lifted Polyps (D), Dyed Resection Margins (E), Out of Patient images (F), Normal Pylorus (G), Stool Inclusions (H), Stool Plenty (I), Blurry Nothing of value (J), Polyps (K), Normal Cecum (L), Colon Clear (M), Retroflex Rectum (N), Retroflex Stomach (O) and Instruments (P).

have an architecture that allows easy extension and widening of the system. To achieve this, we put especial focus on achieving outstanding processing speed without sacrificing high detection accuracy.

From the speed optimization point of view, our system consists of three main parts. The first is a feature extraction module. It is responsible for handling input data, e.g., videos, images and sensor data, and extracting and providing corresponding features extracted from such the data. The most time-consuming aspect here is the extraction of information from the video frames and images. The second part comprises the analysis and decision making algorithms that implement disease detection and localization functions. The last part is the visualization subsystem. It presents the output of the real-time analysis to the endoscopist. The most challenging aspect here is that the visualization should not introduce any delays, which would make the system unsuitable for live examinations.

In order to create the proper optimization strategy we did the preliminary analysis of these three main system parts, resulting in the following optimization steps. The visualization subsystem is implemented using the modern UI handling frameworks and SDKs, and it already utilizes the benefits of the available hardware accelerated I/O and graphics drawing. Additional hardware-oriented optimization of the visualization subsystem is an installation-specific task and should be performed for each specific hospital environment and medical hardware used, thus we consider it to be outside of this research scope. Next, the feature-based decision-making algorithm for the detection subsystem implements already well-optimized classification algorithms efficiently executed on modern CPUs. In the same way, the localization subsystem was implemented with heterogeneous resource utilization in mind from the very beginning, and it did not require deep optimization until we add support for more complex lesion localizers in our system. Finally, we realized that the most time-consuming computation part in our system is the feature extraction module. To achieve mass-screening capabilities and multi-disease detection, the feature extraction architecture had to be improved. We chose to do this by applying heterogeneous processing elements using GPUs.

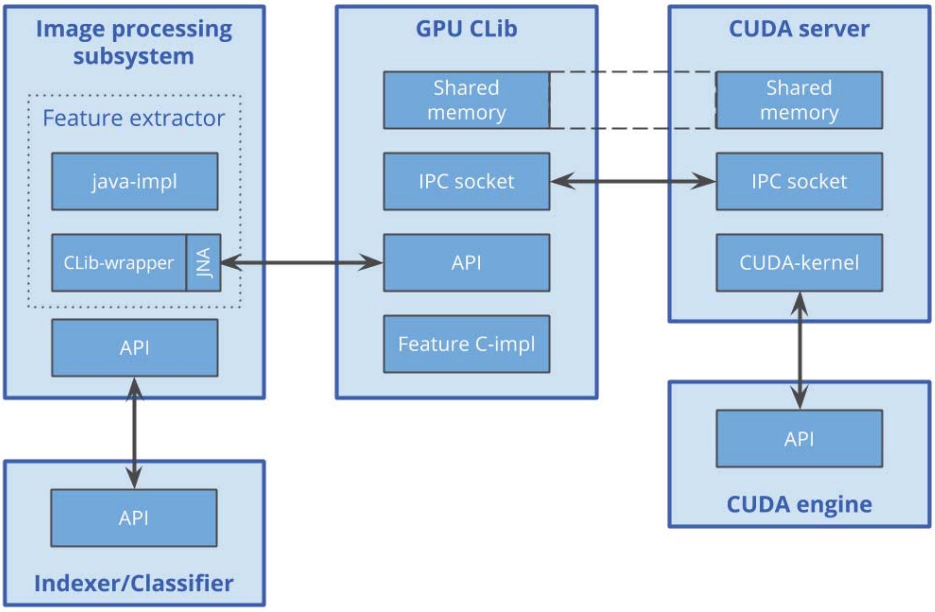


Figure 3.31: The main processing application consisting of the indexing and classification parts uses the GPU-accelerated image processing subsystem. This subsystem provides feature extraction and image filtering algorithms. The most compute-intensive procedures are executed on a stand-alone CUDA-enabled processing server. The interaction between application and server is done via a GPU CLib shared library, which is responsible for maintaining connections and streaming data to and from the CUDA-server.

3.6.3.1 Heterogeneous Architecture

To improve the performance of our feature extraction subsystem, we re-implemented the most compute-intensive parts in CUDA. CUDA is a commonly used GPU processing framework for nVidia graphic cards. We designed the new feature extraction architecture with a heterogeneous processing module as depicted in figure 3.31.

We implemented GPU-accelerated extraction for a number of features (JCD, which includes FCTH and CEDD, and Tamura) for feature-descriptor extraction, as well as for a number of feature-extraction-related procedures, e.g., color space conversion, image resizing and pre-filtering.

In our architecture, as it is shown in figure 3.31, a main processing application interacts with a modular image-processing subsystem. Both are implemented in Java. The image-processing subsystem uses a multi-threaded architecture to handle multiple image processing and feature extraction requests at the same time. All compute-intensive functions are implemented in Java to be able to compare performance with the heterogeneous implementation, which is transparently accessible from Java code through a GPU CLib wrapper. The JNA API is used to access the GPU CLib API directly from the image processing subsystem. The GPU CLib is implemented in C++ as a Linux shared library that connects to a stand-alone processing server and pipes data streams for handling by CUDA implementations. Shared memory is used to avoid the performance penalty of data copying. Local UNIX sockets are used to send requests and receive status responses from the CUDA server because they can be integrated asynchronously on the JNI side than shared-memory semaphores. The CUDA server is implemented in C++ and

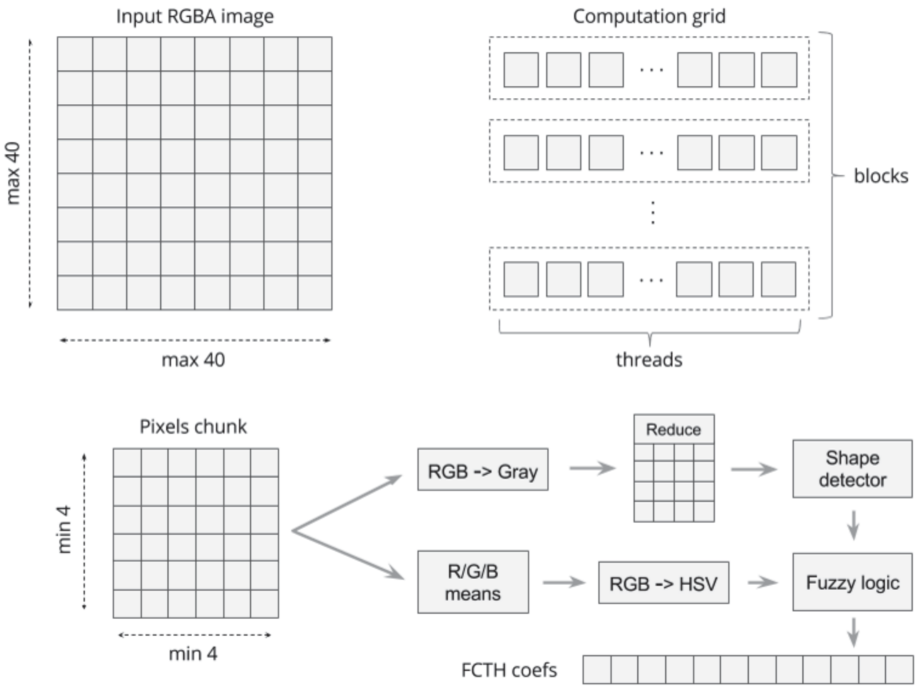


Figure 3.32: GPU-acceleration is used to extract various features from input frames. The figure shows an example of our FCTH feature implementation. The input frame is split into a number of non-overlapping blocks. Each of them is processed separately by two GPU-threads. The main processing steps include color space conversion, size reduction, shape detection and fuzzy logic computations.

uses CUDA SDK to perform computations on GPU. The CUDA server and all heterogeneous-support subsystems are built with distributed processing in mind, and can easily be extended with multiple CUDA servers running locally or on several remote servers.

The processing server can be extended with new feature extractors and advanced image processing algorithms. It enables the utilization of multi-core CPU and GPU resources. As an example, the structure of the FCTH feature extractor implementation is depicted in figure 3.32. It shows that for image features, all pixel-related calculations are executed on the GPU. In case of the FCTH feature, this includes also the processing of a multi-threaded shape detector and fuzzy logic algorithms.

To achieve better performance, a heterogeneous processing subsystem provides the transparent caching of input and intermediate data, which reduces the CPU-GPU bandwidth usage and eliminates redundant data copy operations during image processing.

3.6.3.2 Processing Speed Evaluation

Non-Optimized Architecture

The performance results of the EIR system with non-optimized multi-core CPU-only architecture are depicted in figure 3.33. For all the tests, we used 3 videos from 3 different endoscopic devices and different resolutions. We used three videos of different frame size that are common to widely used endoscopic equipment. These videos are *wp_4* with $1,920 \times 1,080$, *wp_52* with 856×480 and *np_9* with 712×480 frame size, respectively. We chose these videos to show the performance under the different requirements that the system will have to face when in practical

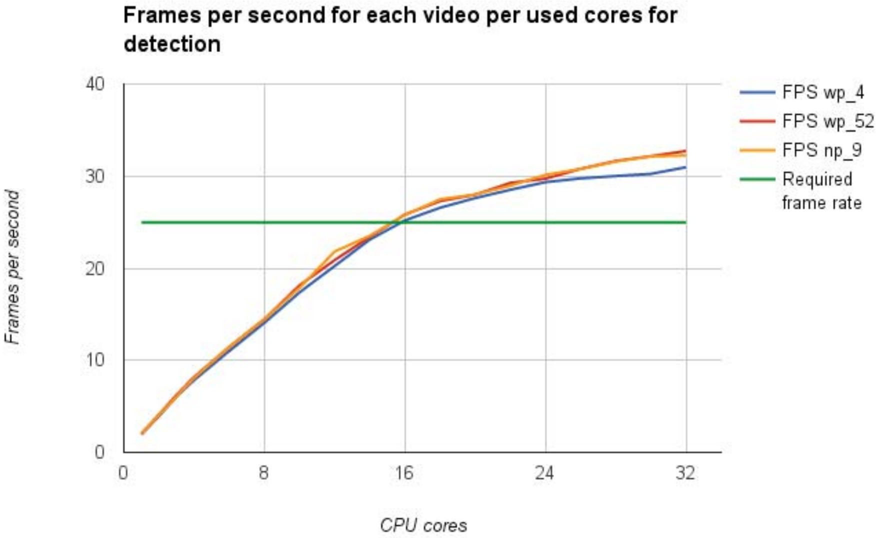


Figure 3.33: The detection performs efficiently and the required frame rate is reached with 12 GB of memory and 16 CPU cores used in parallel on cluster-based computation platform without utilizing heterogeneous architecture.

use. The computer used was a Linux server with 32 AMD CPUs and 128 GB memory. The figures show, that the non-optimized system was able to reach real-time performance for full HD videos using a minimum of 16 CPU cores and at least 12 GB of memory. This has the huge disadvantage that real-time speed is only achieved on expensive highly parallelized multi-CPU systems. In terms of memory, tests showed that the system has rather small requirements. This is beneficial, since it means that memory consumption is not a bottleneck to scalability, and that we can keep this question outside of the optimization process for now.

Heterogeneous Optimized Architecture

The videos used to evaluate the system performance have different resolutions. The resolutions are full HD (1920×1080), WVGA1 (856×480), WVGA2 (712×480) and CIF (384×288). They are labeled correspondingly in figures 3.34, 3.35, 3.36 and 3.37. A framerate of 30 frames per second (FPS) was assumed, and consequently, 33.3 milliseconds processing time per frame was considered real-time speed. Our results for the heterogeneous architecture were obtained using a conventional desktop computer with an Intel Core i7 3.20GHz CPU, 8 GB RAM and a GeForce GTX 460 GPU. To be able to compare the basic and improved systems directly, the same Java source code from the basic system was used to collect the evaluation metrics. In the figures, the basic system's results are labelled as Java. The improved system's results with disabled GPU-acceleration are labelled as C. Finally, the improved system's run in the heterogeneous mode with enabled GPU-acceleration is labelled as GPU.

The performance evaluation shows that the non-optimized architecture can process full HD frames using all 8 available CPU cores and up to 4 GB of memory at 6.5 FPS for Java and 13.8 FPS for the C implementations (see figure 3.34) with corresponding frame processing times of 154ms and 72ms, respectively (see figure 3.36). For the smaller frame sizes, real-time speed was reached at 4 CPU cores and 4 GB of memory. The maximum frame rates that were reached were 49 FPS, 51 FPS and 66 FPS for WVGA1, WVGA2 and CIF frame sizes, respectively (see

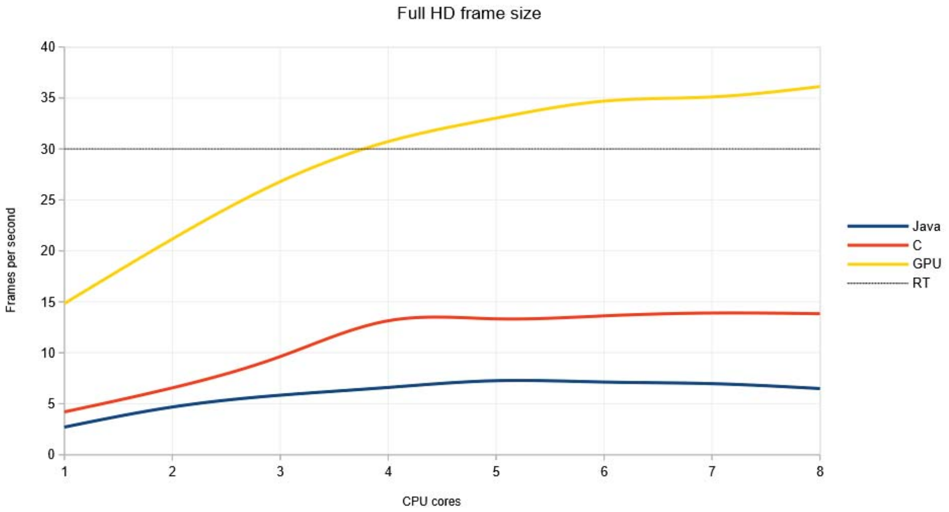


Figure 3.34: The improved GPU-enabled heterogeneous algorithm reaches real-time performance (RT line) with 30 frames per second for full HD (1920 × 1080) videos on a desktop PC using only 4 CPU cores and 5 Gb of memory. The maximum frame rate is around 36 FPS using 8 CPU cores. The Java and C implementations cannot reach real-time performance on the used hardware.

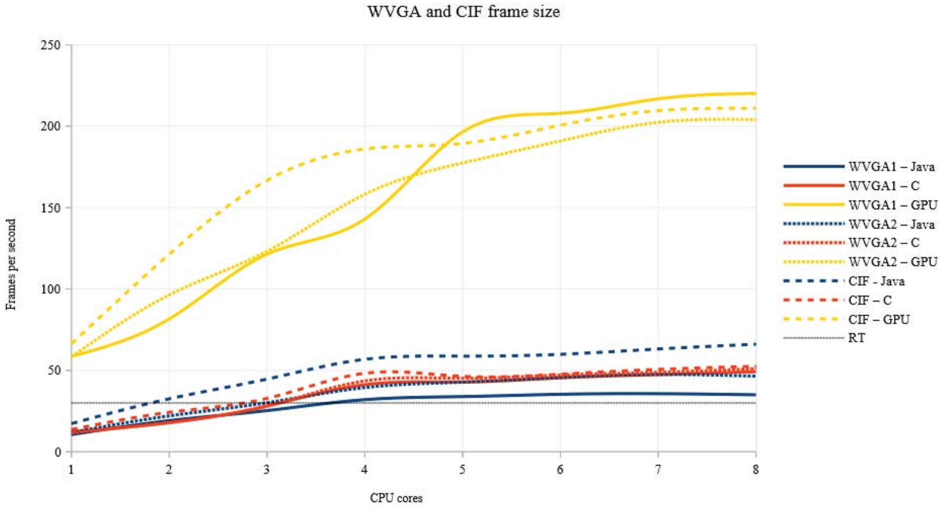


Figure 3.35: The smaller WVGA1 (856 × 480), WVGA2 (712 × 480) and CIF (384 × 288) videos can be processed by the improved GPU-enabled heterogeneous algorithm in real-time using only 1 CPU core. The maximum frame processing rate reaches more than 200 FPS. These results can be improved by putting all feature-related computations on the GPU.

figure 3.35 and figure 3.37).

The evaluation of the improved heterogeneous system shows that the GPU-enabled architecture can easily process full HD frames using only 4 CPU cores (see figure 3.34) and up to 5 Gb of memory with a frame processing time of 32.6ms (see figure 3.36). The maximum frame rate for full HD frames was 36 FPS using all 8 CPU cores. For the smaller frame sizes, the real-time requirements were reached with only 1 CPU core and up to 4.5 GB of memory. The maximum frame rate that we achieved was around 200 FPS (see figure 3.35 and figure 3.37).

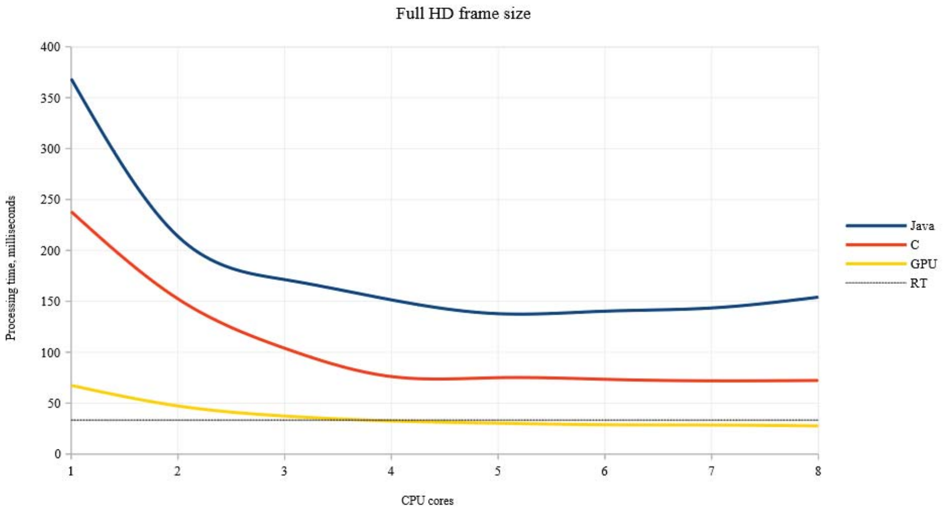


Figure 3.36: The processing time for the GPU-accelerated algorithm decreases slightly with increasing number of used CPU cores for a single full HD frame. This happens due to the CPU-parallel implementation of feature comparison and search algorithms which are not as compute intensive as feature extraction. The Java and C implementations reach the minimum frame processing time with 4 used CPU cores. The reason is that the used CPU has 4 real cores with hyper-threading feature enabled and it cannot handle CPU-intensive calculations efficiently for all 8 (real plus virtual) cores.

The results show clearly that the given hardware system with the basic architecture cannot reach real-time performance for full HD videos even using all available CPU cores, and only for the low-resolution WVGA videos, real-time can be reached. For the improved heterogeneous system, the real-time performance for full HD videos is easily reached using only 4 CPU cores and one outdated GPU. The smaller videos can be processed utilizing only one CPU core plus GPU. Memory size is not a limiting factor and the system can be deployed even on desktop PCs with a general-purpose GPU as an accelerator.

These quantitative results illustrate that using a heterogeneous architecture is key to real-time performance and parallel analysis of videos with different approaches. Furthermore, the improved heterogeneous system has significant over-performance in terms of real-time video processing. This makes it possible to implement more feature extractors, classifiers and many other image processing algorithms to increase the number of detectable diseases by our system while keeping the real-time capability.

3.6.3.3 Distributed Heterogeneous Architecture

The achieved detection performance of 200 frames per seconds is superior with respect to video stream processing time and the ability to provide real-time automatic feedback during live endoscopies. And, even though real-time performance for multiple diseases can be reached by using multiple GPUs in one sufficiently powerful desktop machine, placing such noisy and costly machines in the examination rooms of a hospital is impractical. A more realistic scenario is therefore to have or to use already installed smaller machines in each room, implementing a widely used distributed data processing to use more computation resources whenever more resources are needed. There are many different distributed computation support architectures,

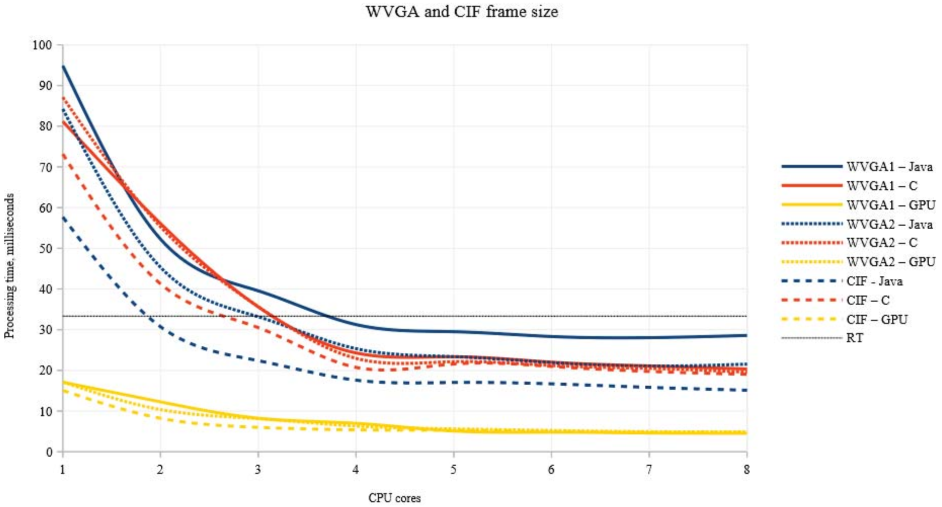


Figure 3.37: For the smaller frame sizes the GPU-accelerated algorithm results in a processing time far below the real-time margin. The minimum is reached with 5 milliseconds using 8 CPU cores. This is a prove for the high system performance and ability to be extended by additional features or to process several video streams at the same time on a conventional desktop PC.

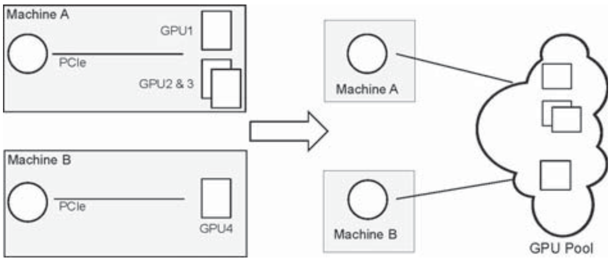


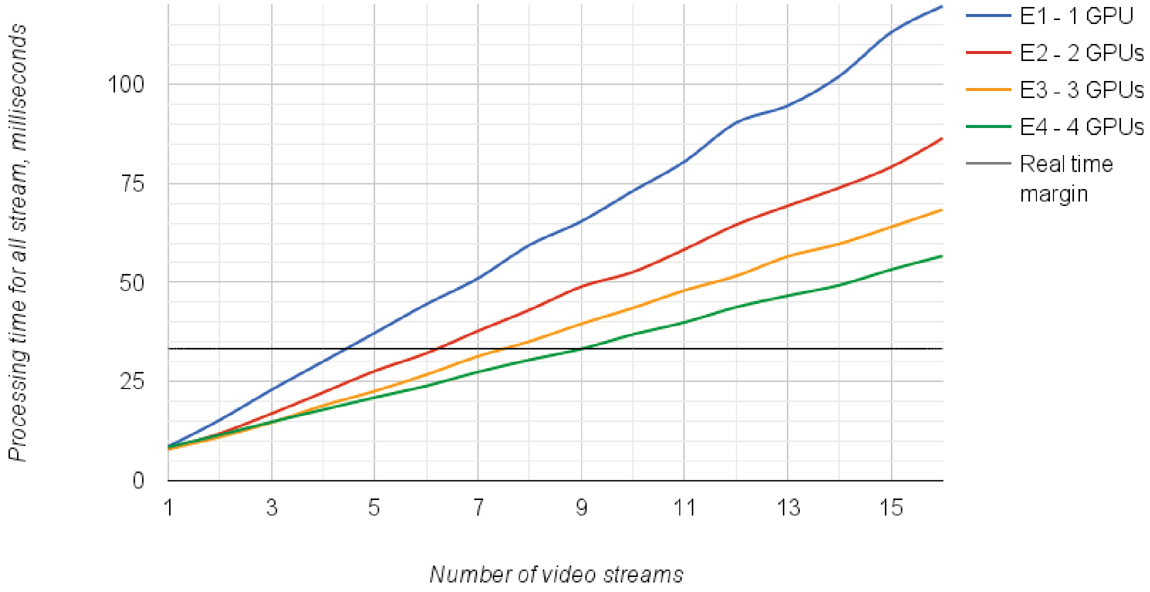
Figure 3.38: Pooling of devices attached in the PCIe network in the experimental setup.

frameworks and SDKs available world-wide, however only few of them are designed with the lowest possible data latency in mind, which is a crucial factor for our real-time-oriented system. Here, the recently developed Device Lending is the best candidate for satisfying our needs to use remote resources locally.

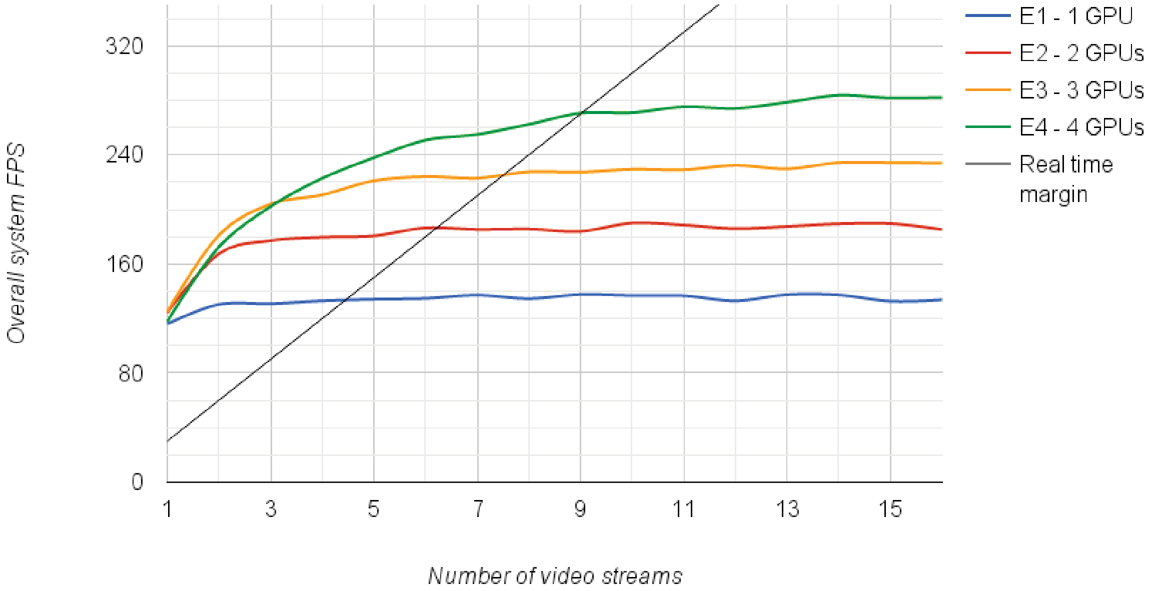
Device Lending is a concept where computers interconnected in a PCI Express [89] network can share devices. It provides transparent, low-latency cross-machine PCIe device sharing (see figure 3.38) without any need to implement application-specific distribution mechanisms or modify native device drivers. The system can allocate and de-allocate additional remote resources, providing dynamic performance management that is able handle workload complexity increases or decreases. It is, therefore, a high-throughput solution can be used for distributed computing, utilizing common hardware already present in all modern computers and requiring little additional interconnection hardware. Device Lending is implemented [73] using Dolphin Interconnect Solutions NTB hardware [11].

For the EIR system, Device Lending enables the combination of multiple GPUs through CUDA's own peer-to-peer communication model, instead of either writing a distributed system, using rCUDA [48] or MPI [86].

To evaluate the performance of the distributed multi-GPU version of our system and also to show that Device Lending in our scenario works as intended, we performed 4 different experiment sets. An overview of the hardware used and the experiments performed can be found in



(a) Frame processing time for several full HD streams in parallel.



(b) Overall system performance for multiple full HD steams in parallel.

Figure 3.39: System performance evaluation in terms of processing time per frame and maximum performance using 4 different configurations described in table 3.21. Each video stream is a full HD video.

Device	Type	E1	E2	E3	E4
GPU1	Nvidia Tesla K40c	*	*	*	*
GPU2	Nvidia Quadro K2200		*	*	*
GPU3	Nvidia GeForce GTX 750			*	*
GPU4	Nvidia Tesla K40c				*

Table 3.21: This table shows the used hardware combinations of the different experiments. GPU 1 to 3 are local GPUs. GPU4 is lend via Device Lending.

table 3.21. For all configurations, we used the same CPU (Intel Core i7-4820K 3.7GHz) and RAM (16GB Quad Channel DDR3). The test setup consists of 2 computers (Machine A and B, see figure 3.38), where the host code of the tests runs on one of them. The second one lends a GPU to it. Experiment E1 uses one local GPU, E2 uses two local GPUs and E3 uses three local GPUs. In E4, we borrowed one GPU from the second computer in addition to three local GPUs. Using these hardware configurations, we performed polyp classification and real-time feedback on the video for up to 16 parallel video streams. All video streams are full HD (1920x1080) videos from colonoscopies. We measured the delay from capturing a video frame to showing the output on the screen. The complete evaluation is shown in figure 3.39.

Figure 3.39(a) shows the performance in terms of processing time per frame for all streams simultaneously. The results reveal that for up to 7 parallel full HD streams, the 3 local GPUs are fast enough. For more than 7 streams, GPU lending is required. The graph shows that the more parallel streams are processed, the better is the performance gain from the borrowed GPU. This is due to the overhead for transferring small amount of data, which hinders Device Lending to reach its full potential. This becomes less important when we have more parallel streams, when Device Lending can indeed improve performance.

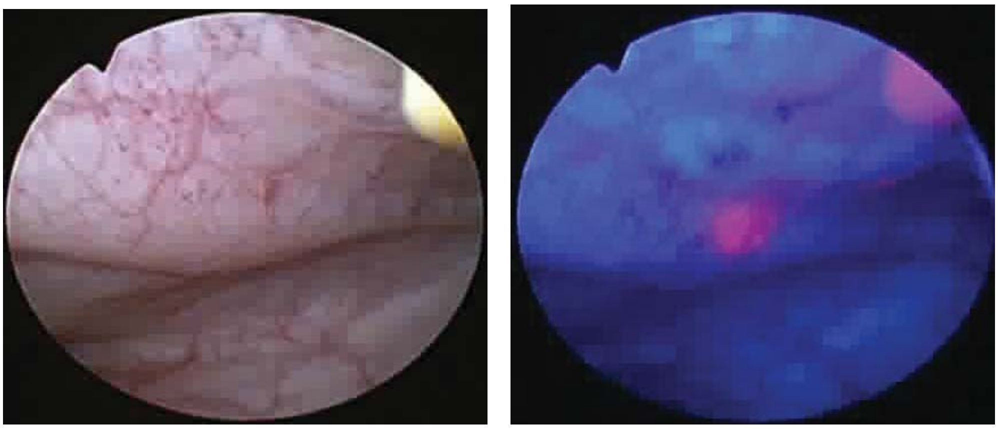
The plot in figure 3.39(b) shows the overall system performance. The maximum overall frames per second we reach when using 4 GPUs at the same time is 30 fps for 9 parallel full HD streams, which is equivalent to 270 fps for a single video stream. Further, this graph shows that the borrowed GPU does not increase the performance for a smaller number of videos, but for 5 and more videos the increase is higher. Thus, the larger amount of data discovers the benefits of the distributed GPU performance boost and, therefore, perfectly fits the multi-auditory examination scenario, while hardware resources are shared within one hospital structure, allowing for mass-screening programs with reduced implementation costs.

3.6.4 System Extensibility Test

For the final system evaluation, we decided to verify our initial claim of easy system extensibility in terms of detected lesions and findings. To perform this, we tested the flexibility of our system using the medical challenges from different application areas that are not directly related to GI tract data analysis, namely bladder cancer cells detection and localization, and spermatozoon localization and segmentation. Both of this two challenges require precise image analysis and introduce additional challenges for the analysis algorithms due to their localization and segmentation nature.

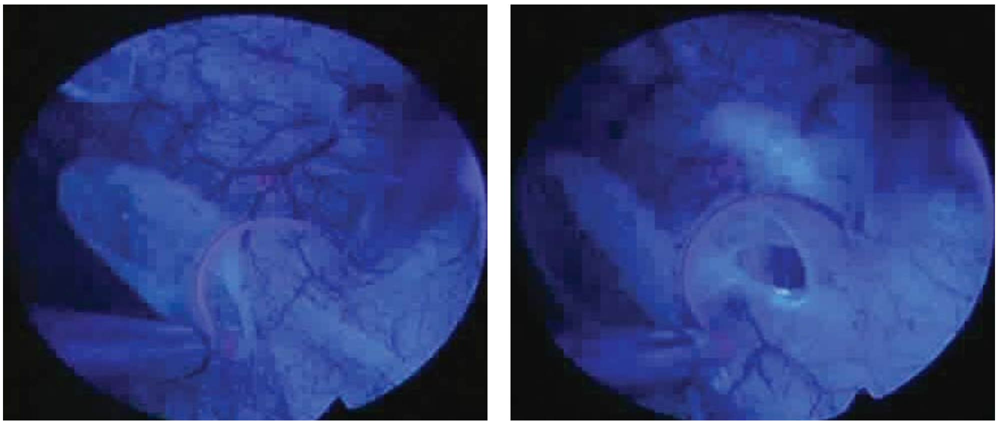
3.6.4.1 Bladder Cancer Cells Detection and Localization

Bladder cancer is the fourth most common cancer and the eighth most common cause of cancer-related mortality in men from the United States [124]. In 2016, roughly 79,030 new cases were diagnosed including 4.6% of all new cancer cases, and 16,870 deaths in the USA were recorded,equating to 2.8% of all cancer deaths [124]. Therefore, initial-stage discovery of bladder cancer is important to reduce risk. The current standard for diagnosis is white-light cystoscopy (WLC) and urine cytology. Complete visualization of the entire bladder and resection of all visible tumors is recommended as a gold treatment standard [36]. Despite its efficiency, the main limitation of WLC is difficulty in identifying all, especially small, areas of malignancy. Current data shows that insufficient detection quality may lead to recurrence of the disease [60]. In contrast, modern blue light cystoscopy (BLC), which is implemented using hexaminolevulinate (named HAL, Cysview or Hexvix) is the most validated technique used today to improve tumor detection. Several prospective trials have shown that HAL-assisted BLC significantly improves the detection of tumors [60]. HAL was approved in EU and US for the



(a) WLC image of an bladder wall area.

(b) BLC image of the same bladder wall area shows a clearly visible tumor cells cluster.



(c) BLC image depicts less visible tumor cells clusters partially be hidden by the interference with blood vessels.

(d) BLC image depicts badly visible tumor cells cluster partially obscured by the resection-remaining tissue.

Figure 3.40: The examples of WLC (a) and BLC (b) frame of our dataset used for the experimental evaluation of the EIR system flexibility and extendability. Images (a) and (b) contain the instrument tip visible in the image top-right corner. Tumor cells clusters are colored by pink color and located in the middle (b), in the middle and top-center (c), and around of the middle (d) of the images.

detection of non–muscle-invasive papillary cancer in patients with suspected bladder lesions. Still, the BLC detection method suffers from limitations in terms of patient population coverage and high miss-rate for small-tumor-cell groups, resulting in around 32% recurrent cancer cases for the BLC-guided examinations [147].

Despite a number of well-developed BLC diagnostic equipment [51], there is a lack of a complete computer-aided bladder cancer cell detection systems. Thus, we selected this use-case as a problem area for verification of our detection and localization subsystems' flexibility and extensibility properties. We adapted our EIR system and in order to provide bladder cancer cell detection and highlighting functionality. To achieve this, we acquired a sample BLC-captured dataset from a Norwegian hospital. The obtained a dataset containing $6,841$ WLC and $7,310$ BLC unannotated and anonymized frames (see figure 3.40 for the example images). The size and variety of our sample dataset does not matter because the goal of this trial with the EIR system is to prove the concept and EIR system flexibility, and not to perform full system training and evaluation. In the following trial run, we used only BLC frames split on the training and test sets. For the training set, we randomly selected 10 BLC images and manually annotated them,

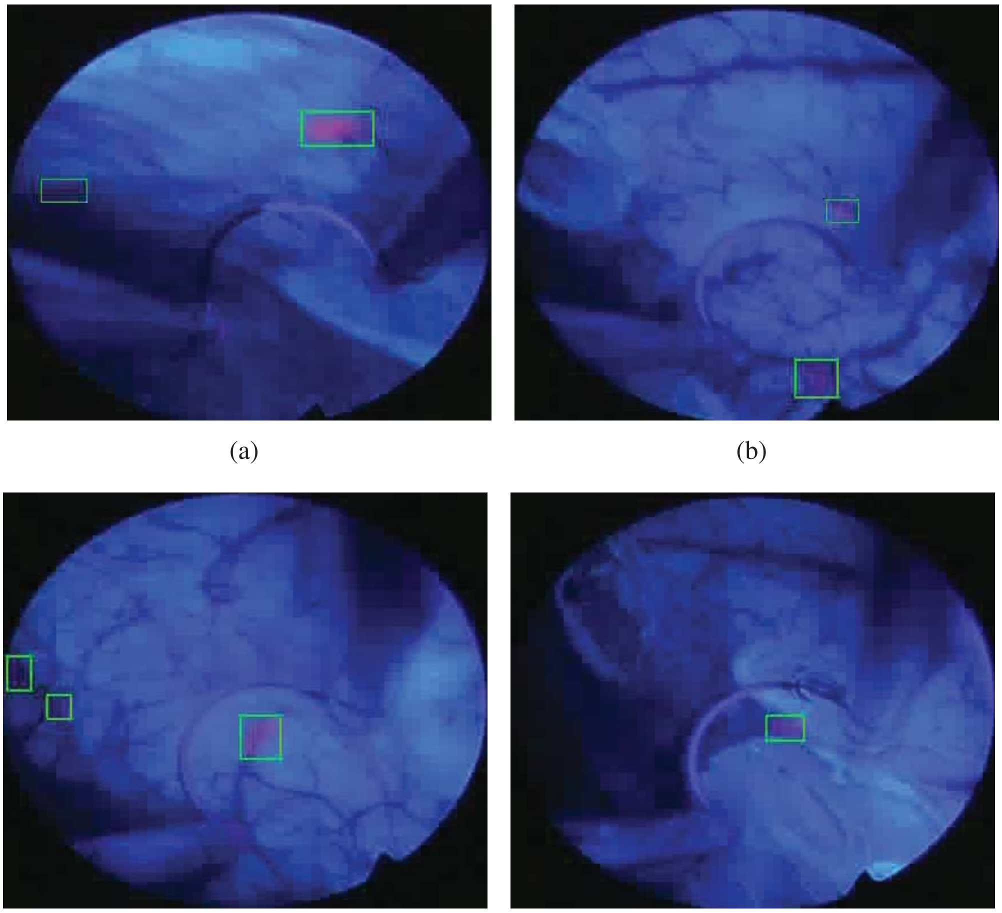


Figure 3.41: The examples of the localized clusters of the bladder cancer cells. The green boxes in the images mark the successfully recognized tumors' locations including ones on the side of the field of view (c), bedly visible in the dark areas (a), located on the blood vessels (b) and partially covered by the tissue (d). One tiny group of cells is missed (e, top-center) probably because of bad input image quality caused by strong video encoding. Constantly visible similarly colored not detected objects are the standard instrument tips.

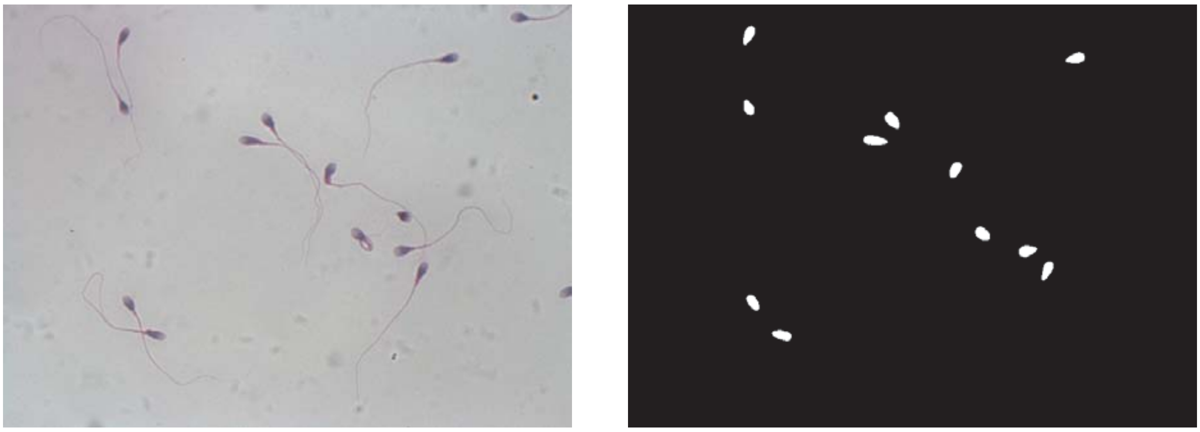
marking the areas showing the tumor cells. Using such a tiny training set allowed us to also test how our EIR system can deal with a new problem area with few annotated data samples, which is especially important for rare but still dangerous diseases.

Using the manually annotated training set, we performed the training of our GAN-based detection and localization approach. The bladder tumor cells have different color and texture properties compared to GI-tract angiectasia lesions, but from the detection and localization point of view, they are similar-looking objects, thus we decided not to perform any fine-tuning on the network or augmentation parameters and used the EIR system as it is. After training, we processed all the training data with the trained model and performed visual performance estimation. The sample detection and localization results are shown in figure 3.41. Without a properly annotated test dataset, it was not possible to evaluate the performance, but the manual inspection of the generated tumor localization boxes confirmed the high quality of the cancer cell cluster marking. The algorithm was able to correctly localize not only clearly visible malignant cell clusters, but also successfully identified clusters that are partially hidden, reside in darkness, are located on the side of the field-of-view or blurred because of camera motions. Moreover, these promising results were obtained using low-quality video footage. With better image quality, we can expect a bladder tumor detection and localization performance as outstanding as we achieved for angiectasia lesion.

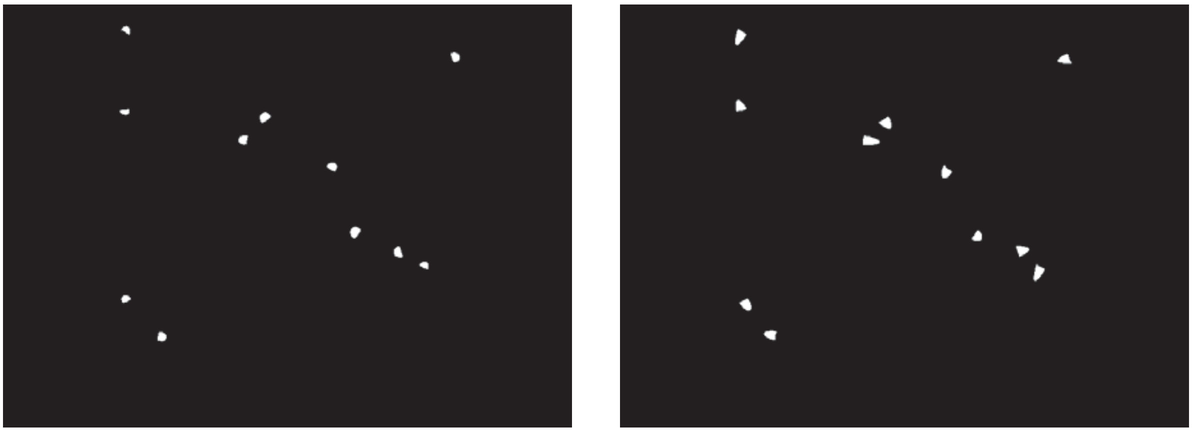
3.6.4.2 Spermatozoon Localization and Segmentation

Semen analysis is routinely used in the fertilization field of applied medicine to evaluate the male partner in infertile couples and to assess the reproductive toxicity of environmental or therapeutic agents [56]. One of the most important factors of sperm quality that can be directly measured is spermatozoons' motility. The estimation of sperm motion parameters using computer-aided sperm analysis improves the objectivity, precision, and reproducibility of the values measured and quantitative motion parameters, such as sperm velocity, and characteristics of track direction can be determined. Computer-aided sperm analysis (CASA) variables, such as progressive motility, linearity, curvilinear velocity, and average path velocity, may serve as prognostic indicators for the fertilization potential of sperm. The measurement of quantitative motility and sperm concentration using CASA is of significant clinical value in predicting the ability of a given ejaculate to achieve successful fertilization and pregnancy in vivo without interventions [47]. Thus, the main goal of a CASA system development is to provide a new methods for automatically detecting and predicting different aspects of human fertility including predicting the motility and morphology of sperms that will lead to a significant reduction of a doctor's workload. Motility and morphology are key attributes [47] for determining the quality of a given sperm sample. Motility is estimated by the individual movement of each spermatozoon, while morphology investigates the shape and form of the sperm cells. Beside the overall sperm quality assessment, another potential use-case is tracking individual spermatozoons in real-time. Thus, the main goals of this preliminary evaluation is to test if the EIR system can be used for this use-case out-of-the-box without any significant modifications.

The crucial factor to the motility and morphology attribute measurement is the spermatozoon localization and morphological segmentation. For the morphology analysis, in the context of semen, doctors often examine the three parts that make up a spermatozoon. These include the



(a) Input microscopic image.
(b) Ground truth mask for heads.

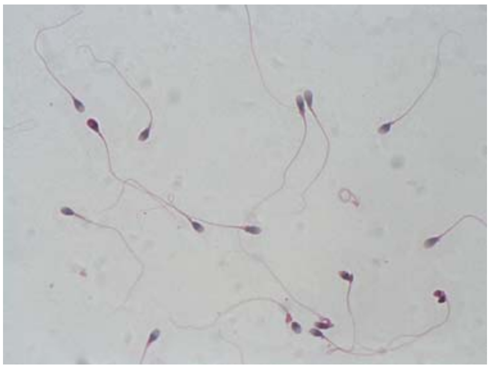


(c) Ground truth mask for acrosomes.
(d) Ground truth mask for nucleuses.

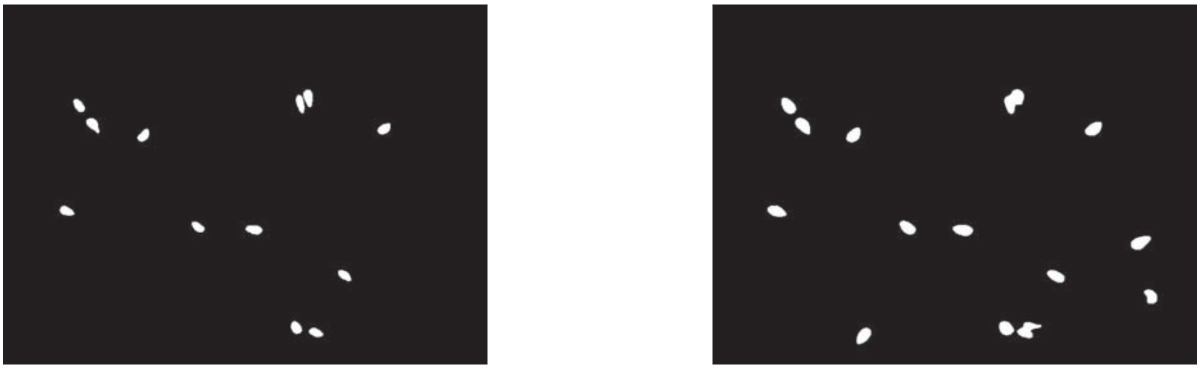
Figure 3.42: The example images of the spermatozoon localization and segmentation dataset used for the experimental evaluation of the EIR system with the different use-case study. First image (a) depicts the source microscopic image in RGB color space. Three other images (b-d) represent the ground truth masks for the different morphological parts of the spermatozoons shown on the image (a).

head (a whole spermatozoon body without a tail), the acrosome (a front-piece of spermatozoon head) and the nucleus (a middle part of a whole spermatozoon in between a acrosome and a tail, rear-piece of spermatozoon head). For the motility estimation, frame-by-frame tracking of the spermatozoons' heads and acrosome positions gives enough information for the travel direction and speed estimation. To the best of our knowledge, there is no a complete CASA system that can solve this semen analysis tasks at once. Collecting a sperm-related dataset and applying our developed detection and localization approaches is our first step in the direction of CASA system development.

The dataset we used in the spermatozoon localization and segmentation experiment consists of 20 RGB frames recorded during a normal sperm microscopy procedure (see figure 3.42 for an example). Each microscopic frame comes along with three different ground truth masks for the different morphological parts of spermatozoon: head, acrosome and nucleus. We split the whole dataset half-and-half into training and testing sets. Than we trained our GAN-based detection, localization and segmentation approach using the corresponding training data. In total, we trained three different independent models for head, acrosome and nucleus. To test the extensibility of our approach, we did not alter any of the training and processing parameters of our

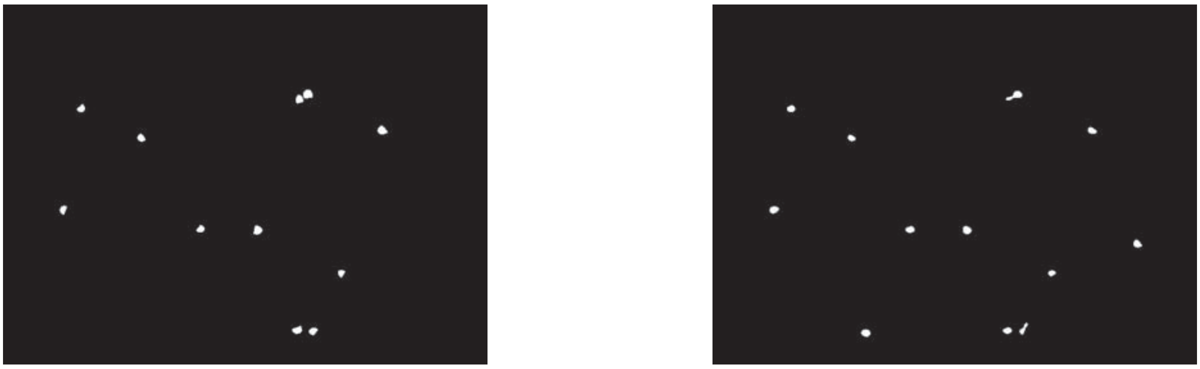


(a) Input microscopic image.



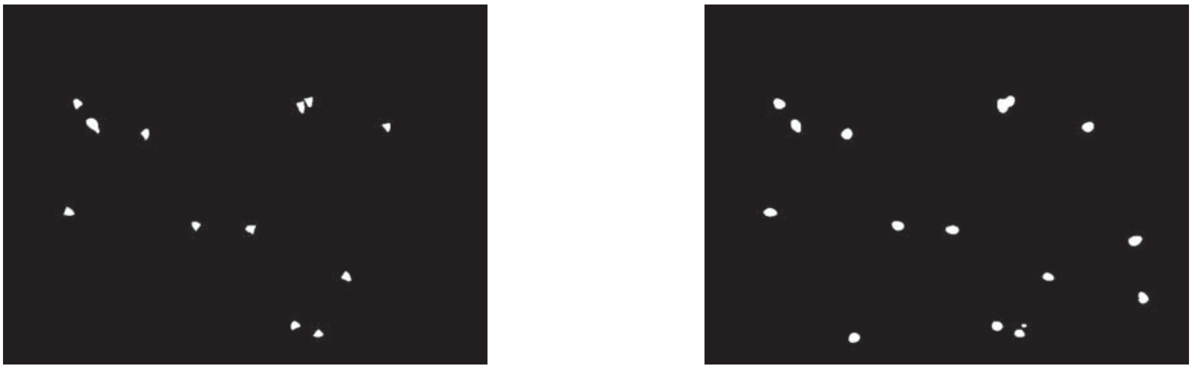
(b) Ground truth mask for heads.

(c) Head segmentation results.



(d) Ground truth mask for acrosomes.

(e) Acrosome segmentation results.



(f) Ground truth mask for nucleuses.

(g) Nucleus segmentation results.

Figure 3.43: The comparison of the ground truth segmentation masks with the output generated segmentation masks of the different morphological parts of the spermatozoons.

networks and used those that were successful for polyp detection and localization. Next, using the trained models, we processed the test dataset in order to generate segmentation masks for the corresponding spermatozoon parts. The example of the three model runs for head, acrosome and nucleus are depicted in figure 3.43. We have not computed any of performance numbers because of a very limited dataset size and because of the incompleteness of the annotation data. For example, figure 3.42(a) shows a clearly visible spermatozoon in the top-left corner, while ground truth data does not have any corresponding markings for this particular object. The same can be observed for two spermatozoons in figure 3.43(a) in the bottom-right corner. Counting the fact that our approach was able to correctly recognize the spermatozoons in this cases, we can state that our approach works well for this use-case. And, as one can see, the generated segmentation masks fit nicely the ground truth, confirming that our polyp-oriented approach can be efficiently retrained to process not only new classes of human tissue lesions, but also perform well for data from different use-case.

3.7 Summary

In this section, we presented our approach for a holistic and complete medical multimedia system called DeepEIR targeted to detect, localize and highlight diseases in the GI tract. The DeepEIR system consists of the complete pipeline from annotation, over detection, localization, segmentation and automatic analysis to visualization. We demonstrated that all parts of the system are important by themselves, and together form a complete system.

We started the DeepEIR system development with the collection of data, training and evaluation of the system performance. We investigated the privacy and legal issues and made agreements with the partner hospitals in Norway to obtain and publish the medial data. We created and published [94, 95, 100] three multi-disease multi-class datasets as open-access resources. There have already received a lot of attention in the research community. We started the medical data analysis competition within the bigger multimedia evaluation benchmark workshop, and we are running it already for three years in row [61, 100, 119].

The data exploration and annotation subsystem is an essential part of the DeepEIR system, because without properly annotated data, it is not possible to train, verify and validate the whole system and its separate components. Moreover, the annotation subsystem allows us literally to transfer medical knowledge data into the IT domain in order to understand and solve the complex and often unexplored multimedia challenges of the medical field without having a deeply specialized medical background and education. It is a well-known fact that medical experts are always very busy. In our annotation subsystem, we tried to address this by introducing an easy-to-understand and use set of tools for data annotation. We developed several annotation tools for medical experts and performed research on these tools to find ones that are better usable and acceptable for the doctors [98, 120].

Next, we developed several modules for the detection subsystem based on different image processing methodologies. First, we extended our single-class global-feature-based detector [97, 115] with new features and classification algorithms [116, 117]. We also made the search-based classification subsystem open source [90], and contributed to the open-source library LIRE, which is used for global features extraction [80]. Than we extended our global-feature-based detector to multi-class use-cases, which allowed us to perform multi-class evaluation experiments with our newly collected multi-disease and multi-object dataset [95]. As a natural step forward, we designed and implemented deep-learning-based [117] and deep-feature-based [87] single- and multi-class classifiers for the detection subsystem, and evaluated and compared them with global-feature-based classifiers [99]. We demonstrated that our detection system can reach a detection performance comparable with state-of-the-art polyp detection approaches, while providing higher processing speeds and reaching our real-time goals [91, 119].

Then, we designed and developed our own hand-crafted local-feature-based polyp localization approach, which is able to spot polyp locations within video frames using polyp color, texture and shape properties. With this spot localizer and our detection subsystem, we successfully participated in the MICCAI challenge [25] for polyp detection and localization. In this challenge, while competing with the research teams working on the polyp recognition for many years, we managed to reach the middle of the overall score for the detection and localization sub-challenges, and we were the second best participant in the detection latency part [25].

The localization subsystem was further extended with new sub-region-based polyp localization modules, each implemented on top of our deep-learning- and deep-feature-based detectors. Here we used splitting of images into smaller, overlapping sub-images with a subsequent detection and detection-result integration to achieve location-based polyp presence estimation and detection [92]. Finally, we implemented universal GAN-based localization-via-segmentation and detection-via-localization modules, which allowed us to achieve both frame- and pixel-wise high-precision polyp detection and localization [92]. We later extended this approach to bleeding [129] and angiectasia [93] lesions, which resulted in outstanding detection and localization performance, which is to our best knowledge, better than the state-of-the-art in angiectasia detection and localization.

To meet real-time speed for Full HD frames, we investigated performance-related issues and evaluated performance of the complete DeepEIR pipeline on different hardware resources. We showed that not all developed subsystems can be executed within real-time constraints using only CPU resources. Therefore, we implemented, presented and evaluated an improved version of the DeepEIR system, which uses a heterogeneous architecture utilizing GPU-acceleration [101]. Even further, we implemented and evaluated distributed workload processing using Device Lending of remote GPUs [102]. The comprehensive results demonstrate that using of heterogeneous resources is the key to real-time performance, and parallel and distributed analysis of multimedia data is a gateway to massive data analysis, which can enable national-wide screening. The developed resource-sharing approach also enables in-hospital hardware resources re-utilization, which leads to reduced installation costs of computer lesion detection systems [118, 121]. We demonstrated that the improved DeepEIR system reaches the outstanding better-than-real-time processing performance of 300 FPS for Full HD video frames, making it possible to implement massive data processing services or add more preprocessors, global- and deep-feature extractors, classifiers, localizers and complex image analysis and processing algorithms to increase the number of detectable diseases by our system while keeping the real-time capability [117, 118].

For the visualization subsystem, we presented three different solutions that can be used by medical experts. These are an online web-based visualization and search tool [80, 90], a real-time polyps detection and spotting tool [91, 96] and a real-time universal lesion detection and localization software. We evaluated the developed visualization subsystem for the real-world use-cases and set the goals for further improved interaction between doctors and computer-aided support systems [116, 117].

Based on the different datasets, including three of our own, we showed that the DeepEIR system can achieve very good results for polyp and other lesion detection and localization while providing real-time feedback to medical doctors while they are performing colonoscopies [91]. We showed that the detection and localization subsystem can reach and for some use-cases outperform state-of-the-art algorithm performance [96, 117]. The whole system was tested by our collaborating medical doctors and was found promising and ready for clinical prototype development [91, 117]. At the moment, DeepEIR is only tested with visual information, but it is built in a way that it can easily be extended to other multimedia data such as sensor or patient data.

Finally, we stress-tested the DeepEIR system for its flexibility and extensibility by running a short successful trial with diseases from different use-case areas, namely bladder cancer cells detection and spermatozoon localization and segmentation. Additionally, we modified and applied our GAN-based localization module to satellite imagery analysis [13, 16, 122], which allowed us to achieve the best flooding areas segmentation performance in the relevant challenges [14, 15].

Thus, in summary, DeepEIR fulfills the requirements set in section 1.2. It is a significant step towards a clinical-ready medical multimedia system that can really help the medical sector in detection, localization, treatment and prevention of some of the most lethal diseases and their short- and long-term consequences, and directly improve the health care system for the whole human society.

Chapter 4

Conclusion

Researching and developing a holistic multimedia medical-purpose-oriented system that can be used for the GI tract disease detection and localization is a complex and multi-disciplinary task requiring investigations in many different problem areas. The work described in this thesis employs both newly developed and state-of-the-art information processing and analysis methods in order to achieve a superior detection and localization performance for the different lesion and ordinary objects of the human GI tract with an outstanding data processing speed and real-time capabilities.

4.1 Summary and Contributions

In this thesis, we presented our experiences with researching and developing a complete holistic medical multimedia system for GI tract disease detection and localization. To stay in the scope of the thesis, we focused on the use case of GI disease and object detection and localization using videos and images. We aimed and were able to build a system that is flexible, generalizable, adaptable, efficient and accurate. As a result, the most important outcome of this work is the DeepEIR system, which reaches high accuracy for lesion and object detection and localization. DeepEIR is easily expandable with new use-cases and data types, runs in real-time, and at the moment the complete system is being tested by medical experts for real clinical studies and trials.

This thesis contributes to several areas of multimedia research. We contributed by researching and developing a medical multimedia system called DeepEIR including data collection, annotation, detection, localization and visualization tools that demonstrates the potential of multimedia research for the health care system.

We started our research from the deep analysis of human GI tract lesion and abnormalities detection needs. We investigated the medical field challenges, with a special focus on the data acquisition and use. We discovered the existing lesion detection and localization approaches, as well as the existing relevant datasets. We made agreements with the collaborating medical institutions and managed to download fully anonymized data for our research purposes.

We collected, annotated and published several new medical datasets freely available under an open-source licenses for research and educational purposes. We researched and developed an efficient set of generalizable and multi-purpose visual-representation-based methods to process and analyze multimedia data. Further, we improved the implementation of methods to achieve real-time and better processing performance and also contributed by researching how distributed processing can be utilized to achieve real-time performance for medical multimedia workload processing. Moreover, we showed some of the privacy and legal issues related to medical multimedia research, demonstrated why the multimedia community should apply their research in medicine, and illustrated how advanced multimedia technology and methods can be used in the medical field to improve workflows, patient care and, most important, potentially save lives. Next, we implemented a set of tools that can be useful for dataset creation regardless of the application area and made the most recent one open source. We implemented and presented several different prototypes and demos of the whole system and various subsystems, and made the detection part of the system open source. Furthermore, we demonstrated that our system is not limited by the primary goal of GI tract inspection, but flexible enough for other types of objects and applications related to visual information analysis. Finally, we contributed by writing and publishing several research papers about our findings and experiences, which we shared with the multimedia research community. We shared our experience regarding how multimedia researchers can apply their knowledge in the medical field and published the article in the ACM multimedia Brave New Idea track [116]. In addition to the DeepEIR system [25, 26, 61, 62, 63, 64, 80, 87, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 115, 117, 118, 119, 120, 121, 129] and side applications of its subsystems [13, 14, 15, 16, 55, 84, 113, 122], this can be seen as an important contribution of this thesis to the research society.

The work presented in this thesis is a continued and extended research on the broad and complex topic of automated lesion detection in the human GI tract. The basic version of the EIR system was jointly developed by Michael Riegler and Konstantin Pogorelov, the author of this thesis. The basic EIR system was described in Riegler's thesis [112]. The second extended and improved version of the EIR system called DeepEIR is presented in this thesis. Both theses include the description of the background, motivation, problem, related work, algorithms and results obtained by Riegler and Pogorelov. The individual author's contributions are explained in chapter 5 and section 1.6.

All main contributions of the thesis are supported by publications in top tier conferences and journals. The contributions to the objectives defined in section 1.2 of the thesis are:

- **Contributions to the main objective:** We developed DeepEIR (the second version of the EIR system) for automatic detection and in-screen localization of lesions in the GI tract, which is capable of giving real-time visual feedback during live colonoscopies using traditional endoscopic equipment as well as of processing huge amount of data for population mass screening using VCEs. The second version of the system consists of an annotation, a detection, a localization and visualization subsystems. The DeepEIR system has been researched and developed with the help of medical experts in our partner hospitals in Norway, Sweden, USA and Austria. The medical experts helped by giving feedback, explaining their field, testing the system and providing data [101, 102, 117, 118, 121].

  Using the ASU Mayo dataset [134], we showed that the detection subsystem of DeepEIR reaches high performance in terms of accuracy and processing. We can report a sensitivity of almost 98% and a precision of almost 94%. This means that DeepEIR is able to find polyps in almost all cases with high precision. This can help the medical experts to save time and lives [101, 102, 117, 118, 121].

Using the recent public Hospital Clinic of Barcelona dataset [23, 24] and our public datasets [94, 95], we showed that the detection subsystem of DeepEIR can reach high frame-wise classification performance in terms of accuracy with the detection specificity of 94% and accuracy of 90.9%. With the same datasets, the localization subsystem reaches the specificity and accuracy of 98.4% and 94.6%, respectively. The resulting performance of our detection and localization approaches is significantly higher than competing global-feature- and deep-learning-based approaches including the most recent real-time YOLOv2 CNN network [107].

Using the angiecstasia segmentation public dataset [23], we showed that the detection and the localization subsystems of DeepEIR can reach outstanding performance that exceeds clinical requirements (sensitivity and specificity higher than 85%). We achieved a sensitivity of 88% and a specificity of 99.9% for pixel-wise angiectasia localization, and a sensitivity of 98% and a specificity of 100% for frame-wise angiectasia detection.

Moreover, we compared DeepEIR with other systems and participated in a classification challenge where we could show that we outperform or reach at least same performance in accuracy as state-of-the-art methods and that we are leading in terms of processing performance [102, 117, 121].

For each part of the DeepEIR system, we developed working prototypes and demo applications. These prototypes and demo applications have been presented at conferences [17, 102, 117, 121].

For the real-time processing challenge, we showed that DeepEIR can process at least 300 FPS for polyp detection, which is a good indicator that we created a scalable medical multimedia system able to process data in real-time [117]. We researched and implemented different ways of distributed and parallel processing using different architectures to improve the performance of the DeepEIR system. One of the methods that we researched is the distribution of the detection and localization part on graphics processing units (GPUs) [101, 121]. Another method that we researched was to distribute the DeepEIR workloads via Device Lending [72, 102]. Both methods improved the processing performance significantly [72, 102].

We showed the potential of multimedia research in the medical field and showed possible further directions and research topics using the DeepEIR system as an example use case [116].

We contributed to two open source projects: *LIRE*, in the field of content-based image retrieval [80], and *OpenVQ*, on video quality [126]. We also released the global-feature-based detection algorithm of DeepEIR as an open source project called Opensea [90].

Finally and most important for us, we contributed with a medical multimedia system for GI examinations that will in the future help medical doctors to save lives.

- **Contributions to sub-objective 1:** For the annotation subsystem of DeepEIR, together with our partner doctors, we did an extensive research in order to make the process of medical knowledge transfer into our system easy and efficient for the medical experts. We explored and developed semi-supervised and cluster-based annotation tools [90, 98, 120].

For the medical data collection and publishing, we researched the ethical and legal aspects of the medical data use within our research process. We contacted several Norwegian hospitals and established relations with the data storage managing personnel. With the help of our medical-side collaborators, we made the agreements allowing us to extract and use the fully anonymized data from the hospital medical information systems. Using these data, we created two datasets (called Kvasir and Nerthus) and published them online freely accessible for educational and research purposes [94, 95]. We used the published datasets for organizing Medico: The 2018 Multimedia for Medicine Task challenge within MediaEval Benchmarking Initiative for Multimedia Evaluation [61, 100, 119]. The public and the research community accepted our Medico challenge. The independent researchers deeply evaluated the datasets and they are already used widely around the world. We also did our evaluation of the datasets to give the baseline for other researchers [87, 99].

- **Contributions to sub-objective 2:** As a basis for the detection subsystem, we developed a search-based classification algorithm that uses global image features, reaches good classification performance and is very fast at the same time [90]. As a basis for the localization subsystem, we developed a polyp localization algorithm based on hand-crafted local features and global heat map post-processing, that is able to reach a good polyp localization precision with a low false-alert rate [25].

  We researched the problem of bleeding detection for VCE-captured videos and developed the basic bleeding detection and localization algorithm for the DeepEIR system [129].

  We implemented the multi-class global-feature- and deep-learning-based classifiers that are able to handle multiple lesions, landmarks and normal findings of the GI tract for the detection subsystem, researched its efficiency both in terms of accuracy and processing speed and compared it with existing competitors [91, 96]. This formed the basis for the DeepEIR system development into the holistic system that is usable and helpful in the real-world conditions.

  In order to extend the lesion detection capabilities of the DeepEIR system, we researched and developed a GAN-based detection and localization approach for the angiectasia GI tract lesion [93]. Also, inspired by the great success of our angiectasia detection approach, we researched and developed a GAN-based polyp detection and localization approach [92].

  We researched the topic of deep neural networks understanding for better medical image classification and classification understanding [62]. We researched the tradeoffs using binary versus multi-class neural network classification for medical multi-disease detection [26].

  Based on the use cases addressed in the thesis and the DeepEIR system itself, we showed that the global- and local-feature-based algorithms together with deep-learning-based approaches can form a strong basis for a multi-lesion detection system. We showed that local hand-crafted features together with GAN-based approaches can provide a good localization performance for the challenging lesions that are hard to see even for humans.

In total, we proved that the developed algorithms are well suited to be applied to several different use cases that involve image classification and analysis problems [91, 92, 93, 99, 101, 102, 116, 117, 118, 121].

- **Contributions to sub-objective 3:** We researched different types of visualization for the output of the DeepEIR system. We developed the specific HTML visualization output generation application for research and medical experts [90] and its easier-to-use web-based version [121]. We developed an initial visualization approach that is able to visualize all outputs of the DeepEIR system [117], which was later involved in the live system output visualization application [96]. We researched the problems of an automatic reporting and decision reasoning system for deep-learning-based analysis in the medical domain [63, 64]

Apart from the main contributions, we also contributed to other multimedia research relevant topics:

Using our GAN-based approach, we researched and developed an approach to the flooding detection on the satellite images that showed promising results [14, 15, 122] and built a unique system for collecting information and monitoring natural disasters by linking social media with satellite imagery can potentially save lives [13, 16].

We researched how the context (a certain watching situation) influences the quality of experience for users when they are watching videos using watching videos during a flight as a use-case. We hosted a MediaEval benchmark task [97] about this topic and published a dataset [115].

We developed a system for efficient live and on-demand tiled HEVC 360 VR video streaming and researched its performance in real use-case scenarios [55].

We researched and developed the new top-down saliency detection approach driven by visual classification showed promising performance on common saliency detection evaluation datasets [84].

In addition to the above contributions, the author also supervised several master students, organized workshops and was part of program committees or conferences.

In summary, we were able to follow a promising and for the society important path by researching and developing a complete medical multimedia system. During this process, we touched and contributed to several areas of multimedia research (annotation, automatic analysis, processing and visualization). We were also able to establish collaborations with several hospitals, which gave us a lot of insight into the medical field and their problems and needs, but also domain knowledge that is needed for creating a useful system. Thus, this work builds a solid basis for future collaboration and work in the field of medical multimedia systems.

4.2 Future Work

For future work, the EIR system can be improved and extended in several ways with new technologies and methods like long short-term memory (LTSM) deep learning approaches for time-based video sequences analysis, advanced pre-processing of images and videos in order to improve detection and localization accuracy, and including more sources of data such as medical sensor data, patient records and audio input from examination rooms. Another important improvement can be a broader comparison of our system with the existing industrial-grade medical systems for GI tract applications in terms of accuracy and usability.

Further widening of the detection and localization capabilities requires the collection of more training data in the various medical fields. The extension of the datasets that have been collected, annotated and published during this work will allow solving even more challenges and will open new possibilities for future research and experiments. Nevertheless, the annotation process of this data is depending on the medical experts and takes a lot of time and effort, and therefore, the collaboration with medical institutions need to be further developed.

The analytical part of the system can be further extended not only with new detectable and localizable diseases and findings, but also with the 3D spatial position localization of the instrument in the whole GI tract using combined motion and landmark analysis. Here, further improvements are also achievable by implementing the 3D reconstruction of the GI tract. A 3D representation of the GI tract could make it easier to detect and localize diseases, position the instrument precisely, and it would also enable lesion size estimation, which is important information for doctors.

The output of an automatic system like DeepEIR also opens many possibilities for visualization, automated reporting and computer-aided diagnosis application scenarios. The automatically selected most-representative samples can be used to add decision-supporting information to patient records such as images of the found diseases or video clips. Moreover, automatic report creation after the examination could help medical doctors to reduce the amount of time spend on reporting. The saved time could then be used to perform additional examinations.

4.3 Final Remarks

Our future research in medical multimedia systems is financially supported by several projects, successfully applied and funded by the Norwegian research council and Oslo Metropolitan University. Within these projects, four PhD students with computer science background and a joint IT-medical PhD student are working jointly to continue this research and enable full-scale clinical trials. The future plan is to make the medical multimedia data and medical expertise publicly available and introduce a ready-to-use system as a routine medical service. This system will be based on our current version of the DeepEIR system and there are a lot of system research and challenges to tackle, i.e., it has to work unattended, preserve privacy, be fault tolerant and well-logged. We fulfilled all research goals that we specified for this thesis and created a holistic system that can be used as a strong basis for future research and applied implementations, and, most important, has the potential to improve the health-care system and actually save lives.

Chapter 5

Papers and Author's Contributions

General overview and discussion of the authors contributions and how the papers contributed to the objectives defined in section 1.2 for each main paper of the thesis. A diagram that also depicts each papers contributions can be found in figure 1.5.

5.1 Paper I: LIRE - Open Source Visual Information Retrieval

Authors: Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, Nektarios Anagnostopoulos

Abstract: With an annual growth rate of 16.2% of taken photos a year, researchers predict an almost unbelievable number of 4.9 trillion stored images in 2017. Nearly 80% of these photos in 2017 will be taken with mobile phones1. To be able to cope with this immense amount of visual data in a fast and accurate way, a visual information retrieval systems are needed for various domains and applications. Lire, short for Luce- ne Image Retrieval, is a light weight and easy to use Java library for visual information retrieval. It allows developers and researchers to integrate common content based image retrieval approaches in their applications and research projects. Lire supports global and local image features and can cope with millions of images using approximate search and distributing indexes on the cloud. In this demo we present a novel tool called F-search that emphasize the core strengths of Lire: lightness, speed and accuracy.

Author's contributions: Pogorelov developed and evaluated the sample (demo) application built on top of LIRE. This application is used in his thesis as the basis for further annotation and visualization tools development. He contributed to the LIRE library code development and did additional performance measurements regarding the search based algorithm. He contributed to all paper sections.

Published in: ACM Multimedia Systems Conference (MMSys), 2016.

Contributed to: Main Objective, Sub-objective 1

See page: 131

5.2 Paper II: OpenSea - Open Search Based Classification Tool

Authors: Konstantin Pogorelov, Zeno Albisser, Olga Ostroukhova, Mathias Lux, Dag Johansen, Pål Halvorsen, Michael Riegler

Abstract: This paper presents an open-source classification tool for image and video frame classification. The classification takes a search-based approach and relies on global and local image features. It has been shown to work with images as well as videos, and is able to perform the classification of video frames in real-time so that the output can be used while the video is recorded, playing, or streamed. OpenSea has been proven to perform comparable to state-of-the-art methods such as deep learning, at the same time performing much faster in terms of processing speed, and can be therefore seen as an easy to get and hard to beat baseline. We present a detailed description of the software, its installation and use. As a use case, we demonstrate the classification of polyps in colonoscopy videos based on a publicly available dataset. We conduct leave-one-out- cross-validation to show the potential of the software in terms of classification time and accuracy.

Author's contributions: Pogorelov was coordinating the writing and submission process. He was responsible for the classification tool testing under different conditions and datasets within the EIR system development and the other side projects. Pogorelov developed an updated version of the OpenSea tool using the updated LIRE library. He conducted a set of experiments with different own and other publicly available datasets in order to validate the tool and approach in general. He wrote the use-case chapter and contributed to other chapters. He prepared and published the open-source repository with the tool for this paper.

Published in: ACM Multimedia Systems Conference (MMSys), 2018.

Contributed to: Main Objective, Sub-objective 1, Sub-objective 3

See page: 137

5.3 Paper III: Explorative Hyperbolic-Tree-Based Clustering Tool for Unsupervised Knowledge Discovery

Authors: Michael Riegler, Konstantin Pogorelov, Mathias Lux, Pål Halvorsen, Carsten Griwodz

Abstract: Exploring and annotating collections of images without meta-data is a laborious task. Visual analytics and information visualization can help users by providing interfaces for exploration and annotation. In this paper, we show a prototype application that allows users from the medical domain to use feature-based clustering to perform explorative browsing and annotation in an unsupervised manner. For this, we utilize global image feature extraction, different unsupervised clustering algorithms and hyperbolic tree representation. First, the prototype application extracts features from images or video frames, and then, one or multiple features at the same time can be used to perform clustering. The clusters are presented to the users as a hyperbolic tree for visual analysis and annotation.

Author's contributions: Pogorelov developed the demo application and the tree-based representation of the clustering output and the annotation part of it. He contributed to the experiments to evaluate the performance of the clustering approach and evaluated the demo application on the medical data. He coded the fast image tree drawing algorithm and optimized the features extraction and clusterization code. He wrote the prototype and demo description section and also contributed to the text in all other sections and the results of these experiments discussion.

Published in: International Workshop on Content-based Multimedia Indexing (CBMI), 2016.

Contributed to: Main Objective, Sub-objective 3

See page: 145

5.4 Paper IV: ClusterTag: Interactive Visualization, Clustering and Tagging Tool for Big Image Collections

Authors: Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Carsten Griwodz

Abstract: Exploring and annotating collections of images without meta-data is a complex task which requires convenient ways of presenting datasets to a user. Visual analytics and information visualization can help users by providing interfaces, and in this paper, we present an open source application that allows users from any domain to use feature-based clustering of large image collections to perform explorative browsing and annotation. For this, we use various image feature extraction mechanisms, different unsupervised clustering algorithms and hierarchical image collection visualization. The performance of the presented open source software allows users to process and display thousands of images at the same time by utilizing GPU resources and different optimization techniques.

Author's contributions: Pogorelov had the idea for the paper. He had the overall responsibility for writing and wrote most of the text in clustering, visualization and the project description sections and contributed to all other sections. He developed the efficient feature extraction, clusterization, real-time database and high-performance drawing algorithms. Pogorelov developed the whole interactive visualization, clustering and tagging tool and performed all the experiments. He did the tool's extensive performance analysis and developed several real-time-oriented caching and on-fly data processing subsystems.

Published in: ACM International Conference on Multimedia Retrieval (ICMR), 2017.

Contributed to: Main Objective, Sub-objective 1, Sub-objective 3

See page: 151

5.5 Paper V: EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal Endoscopies

Authors: Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Thomas de Lange, Carsten Griwodz, Peter Thelin Schmidt, Sigrun Losada Eskeland, Dag Johansen

Abstract: Analysis of medical videos for detection of abnormalities like lesions and diseases requires both high precision and recall but also real-time processing for live feedback during standard colonoscopies and scalability for massive population based screening, which can be done using a capsular video endoscope. Existing related work in this field does not provide the necessary combination of detection accuracy and performance. In this paper, a multimedia system is presented where the aim is to tackle automatic analysis of videos from the human gastrointestinal (GI) tract. The system includes the whole pipeline from data collection, processing and analysis, to visualization. The system combines filters using machine learning, image recognition and extraction of global and local image features, and it is built in a modular way, so that it can easily be extended. At the same time, it is developed for efficient processing in order to provide real-time feedback to the doctor. Initial experiments show that our system has detection and localisation accuracy at least as good as existing systems, but it stands out in terms of real-time performance and low resource consumption for scalability.

Author's contributions: Pogorelov designed and developed a localization approach and the corresponding subsystem. He performed implementation and speed improvements of the detection, analysis and visualization subsystems. He designed and developed experiments for the localization part of the system and contributed to the experiments for the detection part of the system. Pogorelov conducted experiments on the multi-core server and suggested the use of GPU-enabled computations to increase the processing speed and bring real-time capabilities to the EIR system. He contributed to the writing of all the paper's sections.

Published in: International Workshop on Content-based Multimedia Indexing (CBMI), 2016.

Contributed to: Main Objective, Sub-objective 1, Sub-objective 2, Sub-objective 3

See page: 159

5.6 Paper VI: From Annotation to Computer-Aided Diagnosis: Detailed Evaluation of a Medical Multimedia System

Authors: Michael Riegler, Konstantin Pogorelov, Sigrun L. Eskeland, Peter T. Schmidt, Zeno Albisser, Dag Johansen, Carsten Griwodz, Pål Halvorsen, Thomas de Lange

Abstract: In many hospitals, the potential value of multimedia data collected through routine examinations is not recognized. Also, the availability of the data is limited, as the health care personnel have no direct access to the databases where data is stored. However, medical specialists interact with the multimedia content daily through their everyday work and have an increasing interest in finding ways to use it to facilitate their work-processes. In this paper, we present a multimedia system aiming to tackle automatic analysis of video from gastrointestinal (GI) endoscopy. The proposed system includes the whole pipeline from data collection, processing and analysis, to visualization, and it combines filters using machine learning, image recognition and extraction of global and local image features. We built it in a modular way so we can easily extend it to analyze various abnormalities.We also developed it to be efficient enough to run in real-time. The conducted experimental evaluation proves that the detection and localization accuracy reaches at least as good as existing systems' performance, but it is leading in terms of real-time performance and efficient resource consumption.

Author's contributions: Pogorelov contributed to all the development- and evaluation-related sections of the paper. He designed and developed GPU-accelerated detection subsystem, performed and discussed all the detailed performance evaluation experiments in terms of speed and memory consumption for the detection part. He designed and developed the new localization subsystem and its GPU-accelerated implementation, performed the experiments and discussed the results. Pogorelov designed and developed the initial version of the localization subsystem in order to participate MICCAI challenge on polyp detection and localization, and performed all the challenge-related experiments. He also contributed to the real-world use-case and related work sections.

Submitted to: ACM Journal Transactions on Multimedia (ToMM), 2016.

Contributed to: Main Objective, Sub-objective 1, Sub-objective 2, Sub-objective 3

See page: 167

5.7 Paper VII: Multimedia and Medicine: Teammates for Better Disease Detection and Survival

Authors: Michael Riegler, Mathias Lux, Carsten Griwodz, Concetto Spampinato, Thomas de Lange, Sigrun L. Eskeland, Konstantin Pogorelov, Wallapak Tavanapong, Peter T. Schmidt, Cathal Gurrin, Dag Johansen, Håvard Johansen, Pål Halvorsen

Abstract: Health care has a long history of adopting technology to save lives and improve the quality of living. Visual information is frequently applied for disease detection and assessment, and the established fields of computer vision and medical imaging provide essential tools. It is, however, a misconception that disease detection and assessment are provided exclusively by these fields and that they provide the solution for all challenges. Integration and analysis of data from several sources, real-time processing, and the assessment of usefulness for end-users are core competences of the multimedia community and are required for the successful improvement of health care systems. For the benefit of society, the multimedia community should recognize the challenges of the medical world that they are uniquely qualified to address. We have conducted initial investigations into two use cases surrounding diseases of the gastrointestinal (GI) tract, where the detection of abnormalities provides the largest chance of successful treatment if the initial observation of disease indicators occurs before the patient notices any symptoms. Although such detection is typically provided visually by applying an endoscope, we are facing a multitude of new multimedia challenges that differ between use cases. In real-time assistance for colonoscopy, we combine sensor information about camera position and direction to aid in detecting, investigate means for providing support to doctors in unobtrusive ways, and assist in reporting. In the area of large-scale capsular endoscopy, we investigate questions of scalability, performance and energy efficiency for the recording phase, and combine video summarization and retrieval questions for analysis.

Author's contributions: Pogorelov contributed to the showcase and preliminary results sections writing. He designed and implemented the improved GPU-accelerated implementation of the detection and localization subsystems. He contributed to the complete system design description. Pogorelov was responsible for the real-time requirements fulfillment and discussion in the paper. He conducted the performance-related experiments and wrote experiments description and discussion section of the paper. He also contributed to the use-case discussion, did whole paper proof-reading and addressed reviewers' comments.

Published in: ACM Multimedia Conference (MM), 2017.

Contributed to: Main Objective, Sub-objective 1, Sub-objective 2, Sub-objective 3

See page: 195

5.8 Paper VIII: A Holistic Multimedia System for Gastrointestinal Tract Disease Detection

Authors: Konstantin Pogorelov, Sigrun L. Eskeland, Thomas de Lange, Carsten Griwodz, Kristin R. Randel, Håkon K. Stensland, Duc-Tien Dang-Nguyen, Concetto Spampinato, Dag Johansen, Michael Riegler, Pål Halvorsen

Abstract: Analysis of medical videos for detection of abnormalities and diseases requires both high precision and recall, but also real-time processing for live feedback and scalability for massive screening of entire populations. Existing work on this field does not provide the necessary combination of retrieval accuracy and performance. In this paper, a multimedia system is presented where the aim is to tackle automatic analysis of videos from the human gastrointestinal (GI) tract. The system includes the whole pipeline from data collection, processing and analysis, to visualization. The system combines filters using machine learning, image recognition and extraction of global and local image features. Furthermore, it is built in a modular way so that it can easily be extended. At the same time, it is developed for efficient processing in order to provide real-time feedback to the doctors. Our experimental evaluation proves that our system has detection and localisation accuracy at least as good as existing systems for polyp detection, it is capable of detecting a wider range of diseases, it can analyze video in real-time, and it has a low resource consumption for scalability.

Author's contributions: Pogorelov was the coordinator of the paper and contributed to all parts of the paper. Pogorelov designed and developed the first version of the multi-class classifier for the DeepEIR system. He implemented global-features- and deep-feature-based classification subsystems integrated them into DeepEIR and described in the paper. Pogorelov was deeply involved in multi-class data collection for the new medical dataset together with doctors from Vestre Viken Hospital Trust and Cancer Registry of Norway. He performed most of the experiments for system evaluation section, described and discussed the results. He also wrote most of the text for the real-world use cases section. As a result, the paper got an additional ACM Artifact Available label.

Published in: ACM International Conference on Multimedia System (MMSys), 2017.

Contributed to: Main Objective, Sub-objective 1, Sub-objective 2, Sub-objective 3

See page: 207

5.9 Paper IX: GPU-accelerated Real-time Gastrointestinal Diseases Detection

Authors: Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Thomas de Lange, Peter Thelin Smidt, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen

Abstract: The process of finding diseases and abnormalities during live medical examinations has for a long time depended mostly on the medical personnel with some sort of not optimal computer support. However, computer-based medical systems are currently emerging in domains like endoscopies of the gastrointestinal (GI) tract. In this context, we aim for a system that enable automatic analysis of endoscopy videos, where one use case is live computer assisted endoscopies enabling higher disease and abnormality detection rates. In this paper, a system that tackles live automatic analysis of endoscopy videos is presented with a particular focus on the system's capability to perform realtime feedback. The presented system utilizes different parts of a heterogeneous architectures and can be used for automatically analysis of high definition colonoscopy videos (and a fully automated analysis of video from capsular endoscopy devices like pillsized cameras). We describe our implementation and system performance of a GPU-based processing framework. In summary, the experimental results show real-time stream processing and low resource consumption, at a detection precision and recall level at least as good as existing related work.

Author's contributions: Pogorelov introduced the idea of GPU-assisted acceleration of the different parts of the EIR and DeepEIR systems. He designed and implemented GPU-accelerated image and video processing algorithms for the detection subsystem. He did

C++ and CUDA-based implementations of the most compute-intensive blocks of the system. Pogorelov designed, performed and described the experiments in the heterogeneous computational environment. He contributed to all sections of the paper.

Published in: IEEE Computer Based Multimedia System Symposium (CBMS), 2016.

Contributed to: Main Objective, Sub-objective 2

See page: 221

5.10 Paper X: Efficient Processing of Videos in a Multi-Auditory Environment Using Device Lending of GPUs

Authors: Konstantin Pogorelov, Michael Riegler, Jonas Markussen, Håkon Kvale Stensland, Pål Halvorsen, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange

Abstract: In this paper, we present a demo that utilizes Device Lending via PCI Express (PCIe) in the context of a multi-auditory environment. Device Lending is a transparent, low-latency cross-machine PCIe device sharing mechanism without any the need for implementing application-specific distribution mechanisms. As workload, we use a computer-aided diagnosis system that is used to automatically find polyps and mark them for medical doctors during a colonoscopy. We choose this scenario because one of the main requirements is to perform the analysis in real-time. The demonstration consists of a setup of two computers that demonstrates how Device Lending can be used to improve performance, as well as its effect of providing the performance needed for real-time feedback. We also present a performance evaluation that shows its real-time capabilities of it.

Author's contributions: Pogorelov introduced the idea of using device landing for data processing speed improvement of the detection subsystem. He analyzed the possible utilization of device lending for the system speed-up. Pogorelov designed, developed and described distributed and parallel implementation of the algorithms of the detection subsystem. He created the experimental setup, conducted the experiments and analyzed the results. He also contributed to all the sections writing.

Published in: ACM Multimedia Systems Conference (MMSys), 2016.

Contributed to: Main Objective, Sub-objective 2

See page: 229

5.11 Paper XI: Efficient disease detection in gastrointestinal videos - global features versus neural networks

Authors: Konstantin Pogorelov, Michael Riegler, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Carsten Griwodz, Peter Thelin Schmidt, Pål Halvorsen

Abstract: Analysis of medical videos from the human gastrointestinal (GI) tract for detection and localization of abnormalities like lesions and diseases requires both high precision and recall. Additionally, it is important to support efficient, real-time processing for live feedback during (i) standard colonoscopies and (ii) scalability for massive population-based screening, which we conjecture can be done using a wireless video capsule endoscope (camera-pill). Existing related work in this field does neither provide the necessary combination of accuracy and performance for detecting multiple classes of abnormalities simultaneously nor for particular disease localization tasks. In this paper, a complete end-to-end multimedia system is presented where the aim is to tackle automatic analysis of GI tract videos. The system includes an entire pipeline ranging from data collection, processing and analysis, to visualization. The system combines deep learning neural networks, information retrieval, and analysis of global and local image features in order to implement multi-class classification, detection and localization. Furthermore, it is built in a modular way, so that it can be easily extended to deal with other types of abnormalities. Simultaneously, the system is developed for efficient processing in order to provide real-time feedback to the doctors and for scalability reasons when potentially applied for massive population-based algorithmic screenings in the future. Initial experiments show that our system has multi-class detection accuracy and polyp localization precision at least as good as state-of-the-art systems, and provides additional novelty in terms of real-time performance, low resource consumption and ability to extend with support for new classes of diseases.

Author's contributions: Pogorelov was responsible for the whole paper contents and wrote most of the chapters. He designed, developed and implemented a novel local-feature-based polyp localization algorithm. Pogorelov contributed to the multi-class features- and deep-learning-based classification algorithms for DeepEIR detection subsystem and developed GPU-based features extraction code. He conducted a full set of experiments for this paper and performed the performance evaluation and analysis of all the presented approaches. For the first time for DeepEIR system, Pogorelov performed deep analysis of the localization performance and conducted a localization performance comparison to the modern deep-learning-based object localization approaches. He designed, developed and implemented a real-time live polyps detection and localization software.

Published in: Multimedia Tools and Applications (MTAP), 2017.

Contributed to: Main Objective, Sub-objective 1, Sub-objective 2, Sub-objective 3

See page: 235

5.12 Paper XII: Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection

Authors: Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, Pål Halvorsen

Abstract: Automatic detection of diseases by use of computers is an important, but still unexplored field of research. Such innovations may improve medical practice and refine health care systems all over the world. However, datasets containing medical images are hardly available, making reproducibility and comparison of approaches almost impossible. In this paper, we present Kvasir, a dataset containing images from inside the gastrointestinal (GI) tract. The collection of images are classified into three important anatomical landmarks and three clinically significant findings. In addition, it contains two categories of images related to endoscopic polyp removal. Sorting and annotation of the dataset is performed by medical doctors (experienced endoscopists). In this respect, Kvasir is important for research on both single- and multi-disease computer aided detection. By providing it, we invite and enable multimedia researcher into the medical domain of detection and retrieval.

Author's contributions: Pogorelov contributed to all the chapters. He did related work research and analyzed all the relevant publicly available datasets. He was closely involved in the dataset analysis and annotation. He designed and conducted the set of experiments for the reference multi-class classification evaluation using the algorithms from DeepEIR system. He summarized the experimental results. Pogorelov created a website for the dataset and published the dataset with the detailed description online. As a result, the paper got an additional ACM Artifact Available label.

Published in: ACM Multimedia Systems Conference (MMSys), 2017.

Contributed to: Main Objective, Sub-objective 1

See page: 271

5.13 Paper XIII: Nerthus: A Bowel Preparation Quality Video Dataset

Authors: Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, Pål Halvorsen

Abstract: Bowel preparation (cleansing) is considered to be a key precondition for successful colonoscopy (endoscopic examination of the bowel). The degree of bowel cleansing directly affects the possibility to detect diseases and may influence decisions on screening and follow-up examination intervals. An accurate assessment of bowel preparation quality is therefore important. Despite the use of reliable and validated bowel preparation scales, the grading may vary from one doctor to another. An objective and automated assessment of bowel cleansing would contribute to reduce such inequalities and optimize use of medical resources. This would also be a valuable feature for automatic endoscopy reporting in the future. In this paper, we present Nerthus, a dataset containing videos from inside the gastrointestinal (GI) tract, showing different degrees of bowel cleansing.

By providing this dataset, we invite multimedia researchers to contribute in the medical field by making systems automatically evaluate the quality of bowel cleansing for colonoscopy. Such innovations would probably contribute to improve the medical field of GI endoscopy.

Author's contributions: Pogorelov was responsible for the paper writing and submission. He contributed with the dataset creation and anonymized the data before publication. Pogorelov planned, performed and described the basic classification experiments with the dataset. He wrote data collection, dataset details and performance sections. Pogorelov created a website for the dataset and published the dataset with the detailed description online. Together with Riegler, he also was developing and running the web-based two-phase bowel preparation quality assessment survey. The paper got an additional ACM Artifact Available label.

Published in: ACM Multimedia Systems Conference (MMSys), 2017.

Contributed to: Main Objective, Sub-objective 1

See page: 279

5.14 Paper XIV: Deep Learning and Handcrafted Feature Based Approaches for Automatic Detection of Angiectasia

Authors: Konstantin Pogorelov, Olga Ostroukhova, Andreas Petlund, Pål Halvorsen, Thomas de Lange, Håvard Nygaard Espeland, Tomas Kupka, Carsten Griwodz, Michael Riegler

Abstract: Angiectasia, formerly called angiodysplasia, is one of the most frequent vascular lesions and often the cause of gastrointestinal bleedings. Medical specialists assessing videos or images of examinations reach a detection performance of 16% for the detection of bleeding to 69% for the detection of angiectasia. This shows that automatic detection to support medical experts can be useful. In this paper, we present several machine learning-based approaches for angiectasia detection in wireless video capsule endoscopy frames. In summary, the most promising results for pixel-wise localization and frame-wise detection are obtained by the proposed deep learning method using generative adversarial networks (GANs). Using this approach, we achieve a sensitivity of 88% and specificity of 99.9% for pixel-wise localization, and a sensitivity of 98% and a specificity of 100% for frame-wise detection. Thus, the results demonstrate the capability of using deep learning for automatic angiectasia detection in real clinical settings.

Author's contributions: Pogorelov had the initial idea of the paper. He introduced the idea of the paper. He designed and developed a GAN-based segmentation and detection approach for angiectasia lesion, adding a new lesion segmentation functionality to the Deep-EIR system. He planned and performed a set of experiments providing a comprehensive comparison between the GAN-based and deep- and global-feature-based approaches for angiectasia detection. He did a set of cross-validation experiments proving the localization performance efficiency. Pogorelov also was responsible for the paper writing and contributed to all sections.

Published in: IEEE Biomedical and Health Informatics Conference (BHI), 2018.

Contributed to: Main Objective, Sub-objective 1, Sub-objective 2, Sub-objective 3

See page: 287

5.15 Paper XV: Deep Learning and Hand-crafted Feature Based Approaches for Polyp Detection in Medical Videos

Authors: Konstantin Pogorelov, Olga Ostroukhova, Mattis Jeppsson, Håvard Espeland, Carsten Griwodz, Thomas de Lange, Dag Johansen, Michael Riegler, Pål Halvorsen

Abstract: Video analysis including classification, segmentation or tagging is one of the most challenging but also interesting topics multimedia research currently try to tackle. This is often related to videos from surveillance cameras or social media. In the last years, also medical institutions produce more and more video and image content. Some areas of medical image analysis, like radiology or brain scans, are well covered, but there is a much broader potential of medical multimedia content analysis. For example, in colonoscopy, 20% of polyps are missed or incompletely removed on average. Thus, automatic detection to support medical experts can be useful. In this paper, we present and evaluate several machine learning-based approaches for real-time polyp detection for live colonoscopy. We propose pixel-wise localization and frame-wise detection methods which include both handcrafted and deep learning based approaches. The experimental results demonstrate the capability of analyzing multimedia content in real clinical settings, the optimization of the work flow and better detection rates for medical experts.

Author's contributions: Pogorelov introduced the idea of the paper. He designed and developed a combined GAN-based algorithm suitable for implementation of detection, localization and detection-via-localization approaches for DeepEIR system. He tuned his algorithm for the polyp detection and localization use-case and performed the initial proof-of-concept set of experiments. Further, Pogorelov planned designed and performed a set of experiments for through validation of the approach and a comprehensive comparison to the global-features- and deep-learning-based detection approaches. He created and prepared the datasets were used for the experiments. Pogorelov wrote the methodology and experiments sections were also responsible for the whole paper writing and contributed to the paper's text. As a result, the paper got the Best Paper Award from the 2018 IEEE Computer-Based Medical Systems Symposium.

Published in: IEEE Computer-Based Medical Systems Symposium (CBMS), 2018.

Contributed to: Main Objective, Sub-objective 1, Sub-objective 2, Sub-objective 3

See page: 293

Bibliography

[1] ASU-Mayo Clinic Colonoscopy Video Database. `https://polyp.grand-challenge.org/site/Polyp/AsuMayo/`. [last visited, Jul. 12, 2016].

[2] CVC-ClinicDB. `https://polyp.grand-challenge.org/site/Polyp/`. [last visited, Jul. 12, 2016].

[3] CVC Colon Dataset. `http://mv.cvc.uab.es/projects/colon-qa/cvccolondb`. [last visited, Jul. 12, 2016].

[4] ETIS-Larib Polyp DB. `https://polyp.grand-challenge.org/site/Polyp/EtisLarib/`. [last visited, Jul. 12, 2016].

[5] GastroAtlas. `http://www.gastrointestinalatlas.com/index.html`. [last visited, Jul. 12, 2016].

[6] Gastrointestinal Lesions in Regular Colonoscopy Dataset. `http://www.depeca.uah.es/colonoscopy/`. [last visited, Jul. 12, 2016].

[7] GASTROLAB. `http://www.gastrolab.net/index.htm`. [last visited, Jul. 12, 2016].

[8] KID. `https://is-innovation.eu/kid/`. [last visited, Jul. 12, 2016].

[9] The Atlas of Gastrointestinal Endoscopy. `http://www.endoatlas.com/atlas/`. [last visited, Jul. 12, 2016].

[10] WEO Clinical Endoscopy Atlas. `http://www.endoatlas.org/index.php`. [last visited, Jul. 12, 2016].

[11] Dolphin Interconnect Solution PXH810 NTB Adapter, 2015.

[12] M. Abadi, A. Agarwal, P. Barham, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[13] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, and P. Halvorsen. Social media and satellites. *Multimedia Tools and Applications*, pages 1–39, 2018.

[14] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, and H. Pal. Cnn and gan based satellite and social media data fusion for disaster detection. In *Proc. of the MediaEval 2017 Workshop, Dublin, Ireland*, 2017.

[15] K. Ahmad, K. Pogorelov, M. Riegler, O. Ostroukhova, P. Halvorsen, N. Conci, and R. Dahyot. Automatic detection of passable roads after floods in remote sensed and social media data. *Signal Processing: Image Communication*, 74:110–118, 2019.

[16] K. Ahmad, M. Riegler, K. Pogorelov, N. Conci, P. Halvorsen, and F. De Natale. Jord: a system for collecting information and monitoring natural disasters by linking social media with satellite imagery. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, page 12. ACM, 2017.

[17] Z. Albisser, M. Riegler, P. Halvorsen, J. Zhou, C. Griwodz, I. Balasingham, and C. Gurrin. Expert driven semi-supervised elucidation tool for medical endoscopic videos. In *Proceedings of the 6th ACM Multimedia Systems Conference*, pages 73–76. ACM, 2015.

[18] L. A. Alexandre, J. Casteleiro, and N. Nobreinst. Polyp detection in endoscopic video using svms. In *Knowledge Discovery in Databases: PKDD 2007*, pages 358–365. Springer, 2007.

[19] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[20] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino. Texture-based polyp detection in colonoscopy. In *Bildverarbeitung für die Medizin*, pages 346–350. Springer, 2009.

[21] N. N. Baxter, R. Sutradhar, S. S. Forbes, L. F. Paszat, R. Saskin, and L. Rabeneck. Analysis of administrative data finds endoscopist quality measures associated with post-colonoscopy colorectal cancer. *Gastroenterology*, 140(1):65–72, 2011.

[22] J. Bernal and H. Aymeric. Gastrointestinal Image ANAlysis (GIANA) Angiodysplasia D&L challenge. `https://endovissub2017-giana.grand-challenge.org/home/`. Accessed: 2017-11-20.

[23] J. Bernal and H. Aymeric. Miccai endoscopic vision challenge polyp detection and segmentation. `https://endovissub2017-giana.grand-challenge.org/home/`. Accessed: 2017-12-11.

[24] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[25] J. Bernal, N. Tajkbaksh, F. J. Sánchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, K. Pogorelov, S. Choi, Q. Debard, L. Maier-Hein, S. Speidel, D. Stoyanov, P. Brandao, H. Córdova, C. Sánchez-Montes, S. R. Gurudu, G. Fernández-Esparrach, X. Dray, J. Liang, and A. Histace. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE transactions on medical imaging*, 36(6):1231–1249, 2017.

[26] T. J. D. Berstad, M. Riegler, H. Espeland, T. de Lange, P. H. Smedsrud, K. Pogorelov, H. K. Stensland, and P. Halvorsen. Tradeoffs using binary and multiclass neural network classification for medical multidisease detection. In *2018 IEEE International Symposium on Multimedia (ISM)*, pages 1–8. IEEE, 2018.

[27] B. Bilbao-Osorio, S. Dutta, and B. Lanvin. The global information technology report 2013. In *World Economic Forum*, pages 1–383. Citeseer, 2013.

[28] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[29] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[30] H. Brenner, M. Kloor, and C. P. Pox. Colorectal cancer. *Lancet*, 2014.

[31] A. Buades, B. Coll, and J.-M. Morel. Non-local means denoising. *Image Processing On Line*, 1:208–212, 2011.

[32] M. F. Byrne, N. Chapados, F. Soudan, C. Oertel, M. L. Pérez, R. Kelly, N. Iqbal, F. Chandelier, and D. K. Rex. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut*, 68(1):94–100, 2019.

[33] S. Chaabouni, J. Benois-Pineau, and C. B. Amar. Transfer learning with deep networks for saliency prediction in natural video. In *Proc. of ICIP*, pages 1604–1608, 2016.

[34] S. K. Chambers, X. Meng, P. Youl, J. Aitken, J. Dunn, and P. Baade. A five-year prospective study of quality of life after colorectal cancer. *Quality of Life Research*, 21(9), 2012.

[35] S.-F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):688–695, June 2001.

[36] S. S. Chang, S. A. Boorjian, R. Chou, P. E. Clark, S. Daneshmand, B. R. Konety, R. Pruthi, D. Z. Quale, C. R. Ritch, J. D. Seigne, et al. Diagnosis and treatment of non-muscle invasive bladder cancer: Aua/suo guideline. *The Journal of urology*, 196(4):1021–1029, 2016.

[37] S. A. Chatzichristofis and Y. S. Boutalis. CEDD: Color and edge directivity descriptor. a compact descriptor for image indexing and retrieval. In *Proc. of ICVS*, pages 312–322, May 2008.

[38] S. A. Chatzichristofis and Y. S. Boutalis. FCTH: Fuzzy color and texture histogram a low level feature for accurate image retrieval. In *Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2008*, pages 191–196, Klagenfurt, Austria, May 2008. IEEE.

[39] D.-C. Cheng, W.-C. Ting, Y.-F. Chen, Q. Pu, and X. Jiang. Colorectal polyps detection using texture features and support vector machine. In *Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry*, pages 62–72. Springer, 2008.

[40] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[41] L. Cui, C. Hu, Y. Zou, and M. Q.-H. Meng. Bleeding detection in wireless capsule endoscopy images by support vector classifier. In *The 2010 IEEE International Conference on Information and Automation*, pages 1746–1751. IEEE, 2010.

[42] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[43] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of CVPR*, pages 248–255, 2009.

[45] P. J. Denning, D. E. Comer, D. Gries, M. C. Mulder, A. Tucker, A. J. Turner, and P. R. Young. Computing as a Discipline. *Communications of the ACM*, 32(I):1–11, 1989.

[46] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proc. of ICML*, volume 32, pages 647–655, 2014.

[47] E. T. Donnelly, S. E. Lewis, J. A. McNally, and W. Thompson. In vitro fertilization and pregnancy rates: the influence of sperm motility and morphology on ivf outcome. *Fertility and sterility*, 70(2):305–314, 1998.

[48] J. Duato, A. Pena, F. Silla, R. Mayo, and E. Quintana-Ortí. rCUDA: Reducing the number of GPU-based accelerators in high performance clusters. In *Proc. of HPCS*, pages 224–231, 2010.

[49] A. W. Fitzgibbon, R. B. Fisher, et al. A buyer's guide to conic fitting. *DAI Research paper*, 1996.

[50] I. A. for Research on Cancer. *World Cancer Report 2014 (International Agency for Research on Cancer)*, chapter The global and regional burden of cancer. World Health Organization, 2014.

[51] Y. Fradet, H. B. Grossman, L. Gomella, S. Lerner, M. Cookson, D. Albala, M. J. Droller, and P. B. S. Group. A comparison of hexaminolevulinate fluorescence cystoscopy and white light cystoscopy for the detection of carcinoma in situ in patients with bladder cancer: a phase iii, multicenter study. *The Journal of urology*, 178(1):68–73, 2007.

[52] K. Geetha and C. Rajan. Heuristic classifier for observe accuracy of cancer polyp using video capsule endoscopy. *Asian Pac J Cancer Prev*, 18:1681–8, 2017.

[53] B. Giritharan, X. Yuan, J. Liu, B. Buckles, J. Oh, and S. J. Tang. Bleeding detection from capsule endoscopy videos. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4780–4783. IEEE, 2008.

[54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.

[55] C. Griwodz, M. Jeppsson, H. Espeland, T. Kupka, R. Langseth, A. Petlund, P. Qiaoqiao, C. Xue, K. Pogorelov, M. Riegler, et al. Efficient live and on-demand tiled hevc 360 vr video streaming. In *2018 IEEE International Symposium on Multimedia (ISM)*, pages 81–88. IEEE, 2018.

[56] D. S. Guzick, J. W. Overstreet, P. Factor-Litvak, C. K. Brazil, S. T. Nakajima, C. Coutifaris, S. A. Carson, P. Cisneros, M. P. Steinkampf, J. A. Hill, et al. Sperm morphology, motility, and concentration in fertile and infertile men. *New England Journal of Medicine*, 345(19):1388–1393, 2001.

[57] M. Hafner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vecsei, and F. Wrba. Pit pattern classification using extended local binary patterns. In *Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on*, pages 1–4, Nov 2009.

[58] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[59] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of IEEE CVPR*, pages 770–778, 2016.

[60] G. G. Hermann, K. Mogensen, S. Carlsson, N. Marcussen, and S. Duun. Fluorescence-guided transurethral resection of bladder tumours reduces bladder tumour recurrence due to less residual tumour tissue in t a/t1 patients: a randomized two-centre study. *BJU international*, 108(8b):E297–E303, 2011.

[61] S. Hicks, H. L. Hammer, H. K. Stensland, M. Riegler, P. Halvorsen, T. B. Haugen, J. M. Andersen, O. Witczak, R. Borgli, D.-T. Dang-Nguyen, M. Lux, and K. Pogorelov. Medico: The 2019 Multimedia for Medicine Task. `http://www.multimediaeval.org/mediaeval2019/medico/index.html`. [last visited, May. 1, 2019].

[62] S. Hicks, M. Riegler, K. Pogorelov, K. V. Anonsen, T. de Lange, D. Johansen, M. Jeppsson, K. R. Randel, S. L. Eskeland, and P. Halvorsen. Dissecting deep neural networks for better medical image classification and classification understanding. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 363–368. IEEE, 2018.

[63] S. A. Hicks, S. Eskeland, M. Lux, T. de Lange, K. R. Randel, M. Jeppsson, K. Pogorelov, P. Halvorsen, and M. Riegler. Mimir: an automatic reporting and reasoning system for

deep learning based analysis in the medical domain. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 369–374. ACM, 2018.

[64] S. A. Hicks, K. Pogorelov, T. de Lange, M. Lux, M. Jeppsson, K. R. Randel, S. Eskeland, P. Halvorsen, and M. Riegler. Comprehensible reasoning and automated reporting of medical examinations based on deep learning analysis. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 490–493. ACM, 2018.

[65] J. Huang, S. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *cvpr*, volume 97, page 762. Citeseer, 1997.

[66] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. de Groen. Polyp detection in colonoscopy video using elliptical shape feature. In *Proc. of ICIP*, pages 465–468, Sept 2007.

[67] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain. Wireless capsule endoscopy. *Nature*, 405(6785):417, 2000.

[68] Imagenet. ImageNet Challenge Datasets. http://www.image-net.org/. [last visited, March 06, 2016].

[69] M. F. Kaminski, J. Regula, E. Kraszewska, M. Polkowski, U. Wojciechowska, J. Didkowska, M. Zwierko, M. Rupinski, M. P. Nowacki, and E. Butruk. Quality indicators for colonoscopy and the risk of interval cancer. *New England Journal of Medicine*, 362(19):1795–1803, 2010.

[70] J. Kang and R. Doraiswami. Real-time image processing system for endoscopic applications. In *Proc. of CCECE*, volume 3, pages 1469–1472. IEEE, 2003.

[71] J. Kang and J. Gwak. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*, 2019.

[72] L. B. Kristiansen, J. Markussen, H. K. Stensland, M. Riegler, H. Kohmann, F. Seifert, R. Nordstrøm, C. Griwodz, and P. Halvorsen. Device lending in pci express networks. In *Proceedings of the 26th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, page 10. ACM, 2016.

[73] L. B. Kristiansen, J. Markussen, H. K. Stensland, M. Riegler, H. Kohmann, F. Seifert, R. Nordstrøm, C. Griwodz, and P. Halvorsen. Device lending in PCI Express Networks. In *Proc. of NOSSDAV*, 2016.

[74] B. S. Lewis and P. Swain. Capsule endoscopy in the evaluation of patients with suspected small intestinal bleeding: results of a pilot study. *Gastrointestinal endoscopy*, 56(3):349–353, 2002.

[75] B. Li and M.-H. Meng. Tumor recognition in wireless capsule endoscopy images using textural features and svm-based feature selection. *IEEE Transactions on Information Technology in Biomedicine*, 16(3):323–329, May 2012.

[76] B. Li and M. Q.-H. Meng. Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments. *Computers in biology and medicine*, 39(2):141–147, 2009.

[77] Z. Li, D. Carter, R. Eliakim, W. Zou, H. Wu, Z. Liao, Z. Gong, J. Wang, J. W. Chung, S. Y. Song, et al. The current main types of capsule endoscopy. In *Handbook of Capsule Endoscopy*, pages 5–45. Springer, 2014.

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

[79] M. Lux. LIRE: Open source image retrieval in java. In *Proceedings of the 21st ACM MM*, MM '13, page to appear, New York, NY, USA, 2013. ACM.

[80] M. Lux, M. Riegler, P. Halvorsen, K. Pogorelov, and N. Anagnostopoulos. Lire: Open source visual information retrieval. In *Proceedings of the 7th International Conference on Multimedia Systems*, page 30. ACM, 2016.

[81] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE transactions on medical imaging*, 33(7):1488–1502, 2014.

[82] Y. Mori, S.-e. Kudo, M. Misawa, and K. Mori. Simultaneous detection and characterization of diminutive polyps with the use of artificial intelligence during colonoscopy. *VideoGIE*, 4(1):7–10, 2019.

[83] Y. Mori, S.-e. Kudo, M. Misawa, Y. Saito, H. Ikematsu, K. Hotta, K. Ohtsuka, F. Urushibara, S. Kataoka, Y. Ogawa, et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. *Annals of internal medicine*, 169(6):357–366, 2018.

[84] F. Murabito, C. Spampinato, S. Palazzo, D. Giordano, K. Pogorelov, and M. Riegler. Top-down saliency detection driven by visual classification. *Computer Vision and Image Understanding*, 172:67–76, 2018.

[85] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. V. Le, and A. Y. Ng. On optimization methods for deep learning. In *Proc. of ICML*, pages 265–272, 2011.

[86] NVIDIA Corporation. *Developing a Linux Kernel Module using GPUDirect RDMA*, 2015.

[87] O. Ostroukhova, K. Pogorelov, M. Riegler, D.-T. Dang-Nguyen, and P. Halvorsen. Transfer learning with prioritized classification and training dataset equalization for medical objects detection. In *MediaEval 2018 Workshop, Sophia Antipolis, France*, 2018.

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

[89] PCI-SIG. *PCI Express 3.1 Base Specification*, 2010.

[90] K. Pogorelov, Z. Albisser, O. Ostroukhova, M. Lux, D. Johansen, P. Halvorsen, and M. Riegler. Opensea: Open search based classification tool. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 363–368. ACM, 2018.

[91] K. Pogorelov, S. L. Eskeland, T. de Lange, C. Griwodz, K. R. Randel, H. K. Stensland, D.-T. Dang-Nguyen, C. Spampinato, D. Johansen, M. Riegler, et al. A holistic multimedia system for gastrointestinal tract disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 112–123. ACM, 2017.

[92] K. Pogorelov, O. Ostroukhova, M. Jeppsson, H. Espeland, C. Griwodz, T. de Lange, D. Johansen, M. Riegler, and P. Halvorsen. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 381–386. IEEE, 2018.

[93] K. Pogorelov, O. Ostroukhova, A. Petlund, P. Halvorsen, T. de Lange, H. N. Espeland, T. Kupka, C. Griwodz, and M. Riegler. Deep learning and handcrafted feature based approaches for automatic detection of angiectasia. In *Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on*, pages 365–368. IEEE, 2018.

[94] K. Pogorelov, K. R. Randel, T. de Lange, S. L. Eskeland, C. Griwodz, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt, et al. Nerthus: A bowel preparation quality video dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 170–174. ACM, 2017.

[95] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169. ACM, 2017.

[96] K. Pogorelov, M. Riegler, S. L. Eskeland, T. de Lange, D. Johansen, C. Griwodz, P. T. Schmidt, and P. Halvorsen. Efficient disease detection in gastrointestinal videos - global features versus neural networks. *Multimedia Tools and Applications*, 76(21):22493–22525, 2017.

[97] K. Pogorelov, M. Riegler, P. Halvorsen, and C. Griwodz. Simula@ mediaeval 2016 context of experience task. In *MediaEval*, 2016.

[98] K. Pogorelov, M. Riegler, P. Halvorsen, and C. Griwodz. Clustertag: Interactive visualization, clustering and tagging tool for big image collections. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 112–116. ACM, 2017.

[99] K. Pogorelov, M. Riegler, P. Halvorsen, C. Griwodz, T. de Lange, K. Randel, S. Eskeland, D. Nguyen, D. Tien, O. Ostroukhova, M. Lux, and C. Spampinato. A comparison of deep learning with global features for gastrointestinal disease detection. 2017.

[100] K. Pogorelov, M. Riegler, P. Halvorsen, S. A. Hicks, K. R. Randel, D.-T. Dang-Nguyen, M. Lux, O. Ostroukhova, and T. de Lange. Medico multimedia task at mediaeval 2018. In *MediaEval 2018 Workshop, Sophia Antipolis, France*, 2018.

[101] K. Pogorelov, M. Riegler, P. Halvorsen, P. T. Schmidt, C. Griwodz, D. Johansen, S. L. Eskeland, and T. de Lange. Gpu-accelerated real-time gastrointestinal diseases detection. In *Computer-Based Medical Systems (CBMS), 2016 IEEE 29th International Symposium on*, pages 185–190. IEEE, 2016.

[102] K. Pogorelov, M. Riegler, J. Markussen, H. K. Stensland, P. Halvorsen, C. Griwodz, S. L. Eskeland, and T. de Lange. Efficient processing of videos in a multi-auditory environment using device lending of gpus. In *Proceedings of the 7th International Conference on Multimedia Systems*, page 36. ACM, 2016.

[103] M. Porta. New visualization modes for effective image presentation. *International Journal of Image and Graphics*, 9(01):27–49, 2009.

[104] E. Quintero, C. Hassan, C. Senore, and Y. Saito. Progress and challenges in colorectal cancer screening. *Gastroenterology research and practice*, 2012, 2012.

[105] J. Redmon. Darknet: Open source neural networks in C. http://pjreddie.com/darknet/. [last visited, March 06, 2016].

[106] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv:1506.02640*, 2015.

[107] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[108] D. K. Rex. Rationale for colonoscopy screening and estimated effectiveness in clinical practice. *Gastrointestinal endoscopy clinics of North America*, 12(1):65–75, 2002.

[109] D. K. Rex, J. H. Bond, S. Winawer, T. R. Levin, R. W. Burt, D. A. Johnson, L. M. Kirk, S. Litlin, D. A. Lieberman, J. D. Waye, et al. Quality in the technical performance of colonoscopy and the continuous quality improvement process for colonoscopy: recommendations of the us multi-society task force on colorectal cancer. *The American journal of gastroenterology*, 97(6):1296, 2002.

[110] D. K. Rex, P. S. Schoenfeld, J. Cohen, I. M. Pike, D. G. Adler, M. B. Fennerty, J. G. Lieb, W. G. Park, M. K. Rizk, M. S. Sawhney, N. J. Shaheen, S. Wani, and D. S. Weinberg. Quality indicators for colonoscopy. *American J. of Gastroenterology*, 110(1):72–90, 2015.

[111] J.-F. Rey, R. Lambert, and the ESGE Quality Assurance Committee. Esge recommendations for quality control in gastrointestinal endoscopy: guidelines for image documentation in upper and lower gi endoscopy. *Endoscopy*, 33(10):901–903, 2001.

[112] M. Riegler. *EIR - A Medical Multimedia System for Efficient Computer Aided Diagnosis*. PhD thesis, PhD thesis. University of Oslo, 2017.

[113] M. Riegler, D.-T. Dang-Nguyen, B. Winther, C. Griwodz, K. Pogorelov, and P. Halvorsen. Heimdallr: a dataset for sport analysis. In *Proceedings of the 7th International Conference on Multimedia Systems*, page 47. ACM, 2016.

[114] M. Riegler, M. Larson, M. Lux, and C. Kofler. How ’how’ reflects what’s what: Content-based exploitation of how users frame social images. In *Proc. of ACM MM*, pages 397–406, 2014.

[115] M. Riegler, M. Larson, C. Spampinato, P. Halvorsen, M. Lux, J. Markussen, K. Pogorelov, C. Griwodz, and H. Stensland. Right inflight?: A dataset for exploring the automatic prediction of movies suitable for a watching situation. In *Proceedings of the 7th International Conference on Multimedia Systems*, page 45. ACM, 2016.

[116] M. Riegler, M. Lux, C. Griwodz, C. Spampinato, T. de Lange, S. L. Eskeland, K. Pogorelov, W. Tavanapong, P. T. Schmidt, C. Gurrin, et al. Multimedia and medicine: Teammates for better disease detection and survival. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 968–977. ACM, 2016.

[117] M. Riegler, K. Pogorelov, S. L. Eskeland, P. T. Schmidt, Z. Albisser, D. Johansen, C. Griwodz, P. Halvorsen, and T. D. Lange. From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(3):26, 2017.

[118] M. Riegler, K. Pogorelov, P. Halvorsen, T. de Lange, C. Griwodz, P. T. Schmidt, S. L. Eskeland, and D. Johansen. Eir — efficient computer aided diagnosis framework for gastrointestinal endoscopies. In *Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on*, pages 1–6. IEEE, 2016.

[119] M. Riegler, K. Pogorelov, P. Halvorsen, C. Griwodz, T. Lange, K. Randel, S. Eskeland, D. Nguyen, D. Tien, M. Lux, C. Griwodz, C. Spampinato, and T. de Lange. Multimedia for medicine: the medico task at mediaeval 2017. 2017.

[120] M. Riegler, K. Pogorelov, M. Lux, P. Halvorsen, C. Griwodz, T. de Lange, and S. L. Eskeland. Explorative hyperbolic-tree-based clustering tool for unsupervised knowledge discovery. In *Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on*, pages 1–4. IEEE, 2016.

[121] M. Riegler, K. Pogorelov, J. Markussen, M. Lux, H. K. Stensland, T. de Lange, C. Griwodz, P. Halvorsen, D. Johansen, P. T. Schmidt, et al. Computer aided disease detection system for gastrointestinal examinations. In *Proceedings of the 7th International Conference on Multimedia Systems*, page 29. ACM, 2016.

[122] N. Said, K. Pogorelov, K. Ahmad, M. Riegler, N. Ahmad, O. Ostroukhova, P. Halvorsen, and N. Conci. Deep learning approaches for flood classification and flood aftermath detection. In *MediaEval 2018 Workshop, Sophia Antipolis, France*, 2018.

[123] L. Sharp, L. Tilson, S. Whyte, A. O'Ceilleachair, C. Walsh, C. Usher, P. Tappenden, J. Chilcott, A. Staines, M. Barry, et al. Cost-effectiveness of population-based screening for colorectal cancer: a comparison of guaiac-based faecal occult blood testing, faecal immunochemical testing and flexible sigmoidoscopy. *British journal of cancer*, 106(5):805–816, 2012.

[124] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67(1):7–30, 2017.

[125] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[126] K. Skarseth, H. Bjørlo, P. Halvorsen, M. Riegler, and C. Griwodz. Openvq: A video quality assessment toolkit. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1197–1200. ACM, 2016.

[127] J. Son, S. J. Park, and K.-H. Jung. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. *arXiv preprint arXiv:1706.09318*, 2017.

[128] R. Stewart and M. Andriluka. End-to-end people detection in crowded scenes. *arXiv*, 2015.

[129] S. Suman, F. A. B. Hussin, A. S. Malik, K. Pogorelov, M. Riegler, S. H. Ho, I. Hilmi, and K. L. Goh. Detection and classification of bleeding region in wce images using color feature. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, page 17. ACM, 2017.

[130] M. Sumner, E. Frank, and M. Hall. Speeding up logistic model tree induction. In *European conference on principles of data mining and knowledge discovery*, pages 675–683. Springer, 2005.

[131] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[132] C. Szegedy, V. Vanhoucke, S. Ioffe, et al. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015.

[133] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proc. of IEEE CVPR*, pages 2818–2826, 2016.

[134] N. Tajbakhsh, S. R. Gurudu, and J. Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2016.

[135] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, man, and cybernetics*, 8(6):460–473, 1978.

[136] T. T. Tanimoto. elementary mathematical theory of classification and prediction. 1958.

[137] The New York Times. The $2.7 Trillion Medical Bill. `http://www.nytimes.com/2013/06/02/health/colonoscopies_explain_why_us_leads_the_world_in_health_expenditures.html`. [last visited, Oct. 10, 2016].

[138] The New York Times. The Weird World of Colonoscopy Costs. `http://www.nytimes.com/2013/06/09/opinion/sunday/the_weird_world_of_colonoscopy_costs.html`. [last visited, Aug. 29, 2016].

[139] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012.

[140] G. Urban, P. Tripathi, T. Alkayali, M. Mittal, F. Jalali, W. Karnes, and P. Baldi. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, 155(4):1069–1078, 2018.

[141] R. Valori, J.-F. Rey, W. S. Atkin, M. Bretthauer, C. Senore, G. Hoff, E. J. Kuipers, L. Altenhofen, R. Lambert, and G. Minoli. European guidelines for quality assurance in colorectal cancer screening and diagnosis. first edition – quality assurance in endoscopy in colorectal cancer screening and diagnosis. *Endoscopy*, 44(S03):SE88–SE105, 2012.

[142] B. Van Essen, C. Macaraeg, M. Gokhale, and R. Prenger. Accelerating a random forest classifier: Multi-core, GP-GPU, or FPGA? In *Proc. of FCCM*, pages 232–239, 2012.

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

[144] L. von Karsa, J. Patnick, and N. Segnan. European guidelines for quality assurance in colorectal cancer screening and diagnosis. first edition–executive summary. *Endoscopy*, 44(S 03), 2012.

[145] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy. *Biomedical and Health Informatics, IEEE Journal of*, 18(4):1379–1389, 2014.

[146] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. De Groen. Polyp-alert: Near real-time feedback during colonoscopy. *Computer methods and programs in biomedicine*, 120(3):164–179, 2015.

[147] J. A. Witjes, M. Babjuk, P. Gontero, D. Jacqmin, A. Karl, S. Kruck, P. Mariappan, J. P. Redorta, A. Stenzl, R. Van Velthoven, et al. Clinical and cost effectiveness of hexaminolevulinate-guided blue-light cystoscopy: evidence review and updated expert recommendations. *European urology*, 66(5):863–871, 2014.

[148] World Health Organization - International Agency for Research on Cancer. Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. `http://globocan.iarc.fr/Pages/fact_sheets_population.aspx`. [last visited, Jul. 12, 2016].

[149] K. Zagoris, S. A. Chatzichristofis, N. Papamarkos, and Y. S. Boutalis. Automatic image annotation and retrieval using the joint composite descriptor. In *2010 14th Panhellenic Conference on Informatics*, pages 143–147. IEEE, 2010.

[150] M. Zhou, G. Bao, Y. Geng, B. Alkandari, and X. Li. Polyp detection and radius measurement in small intestine using video capsule endoscopy. In *Proc. of BMEI*, pages 237–241, Oct 2014.

Part II

Research Papers

Paper XI

Efficient disease detection in gastrointestinal videos - global features versus neural networks

Multimed Tools Appl (2017) 76:22493–22525
DOI 10.1007/s11042-017-4989-y

Efficient disease detection in gastrointestinal videos – global features versus neural networks

Konstantin Pogorelov[1] · **Michael Riegler**[1] · **Sigrun Losada Eskeland**[2] · **Thomas de Lange**[2] · **Dag Johansen**[3] · **Carsten Griwodz**[1] · **Peter Thelin Schmidt**[4] · **Pål Halvorsen**[1]

Received: 14 October 2016 / Revised: 29 May 2017 / Accepted: 27 June 2017 /
Published online: 19 July 2017
© The Author(s) 2017. This article is an open access publication

Abstract Analysis of medical videos from the human gastrointestinal (GI) tract for detection and localization of abnormalities like lesions and diseases requires both high precision and recall. Additionally, it is important to support efficient, real-time processing for live feedback during (i) standard colonoscopies and (ii) scalability for massive population-based screening, which we conjecture can be done using a wireless video capsule endoscope (camera-pill). Existing related work in this field does neither provide the necessary

---

✉ Konstantin Pogorelov
konstantin@simula.no

Michael Riegler
michael@simula.no

Sigrun Losada Eskeland
sigesk@vestreviken.no

Thomas de Lange
t.d.lange@medisin.uio.no

Dag Johansen
dag.johansen@uit.no

Carsten Griwodz
griff@simula.no

Peter Thelin Schmidt
peter.thelin-schmidt@karolinska.se

Pål Halvorsen
paalh@ifi.uio.no

[1] Simula Research Laboratory, P.O. Box 134, 1325, Lysaker, Norway

[2] Bærum Hospital, Lysaker, Norway

[3] UiT-The Arctic University of Norway, Lysaker, Norway

[4] Karolinska Institutet, Solna, Sweden

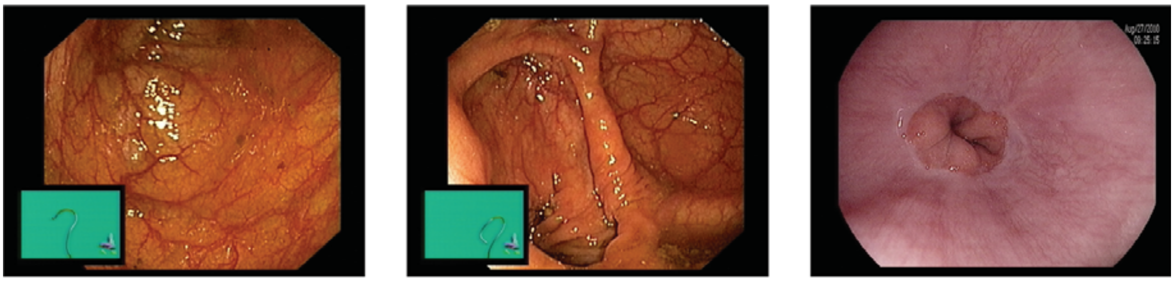
combination of accuracy and performance for detecting multiple classes of abnormalities simultaneously nor for particular disease localization tasks. In this paper, a complete end-to-end multimedia system is presented where the aim is to tackle automatic analysis of GI tract videos. The system includes an entire pipeline ranging from data collection, processing and analysis, to visualization. The system combines deep learning neural networks, information retrieval, and analysis of global and local image features in order to implement multi-class classification, detection and localization. Furthermore, it is built in a modular way, so that it can be easily extended to deal with other types of abnormalities. Simultaneously, the system is developed for efficient processing in order to provide real-time feedback to the doctors and for scalability reasons when potentially applied for massive population-based algorithmic screenings in the future. Initial experiments show that our system has multi-class detection accuracy and polyp localization precision at least as good as state-of-the-art systems, and provides additional novelty in terms of real-time performance, low resource consumption and ability to extend with support for new classes of diseases.

Keywords Medical · Automatic disease detection · Algorithmic screening · Global and local image features · Deep learning neural networks · Information retrieval · Performance evaluation

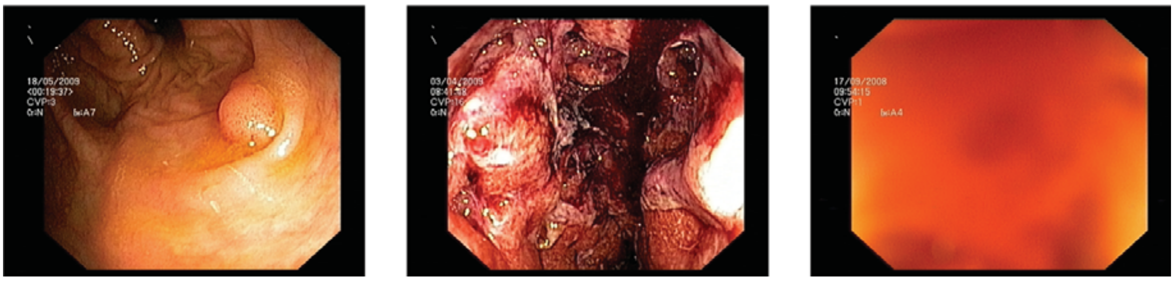
1 Introduction

Rapid development of technologies in areas of sensors, imaging devices and diagnostic methods shifts the paradigm in medical diagnostic from manual analysis by trained doctors to wide usage of automated computer-assisted diagnostic systems. In our research, we are working at the intersection between computer science and pathological medicine, where we target a scalable, real-time, multi-disease detection system for the gastrointestinal (GI) tract. Our aim is to develop both a computer-aided, live analysis system of endoscopy videos and a scalable detection system for population-wide screening using a wireless video capsule endoscope (VCE). This small capsule with one or more image sensors is swallowed and captures videos while it traverses the entire GI tract.

In the context of object detection, localization and tracking in images and videos, a lot of research carried out. Particularly, current systems have been developed to detect general objects from the surrounding world, for example human faces, cars and logos. Our research targets a totally different domain, which is inside the body of a human being. Both the general objects and the GI tract irregularities can have different sizes, shapes, textures, colors and orientations, they can be located anywhere in the frame and also partially be hidden and covered by other objects and obstacle. However, GI tract findings can also have a color, texture and shape properties similar for the different diseases, as well as different for the similar diseases on the various developing stages. The GI findings can be covered by the biological substances, like for example seeds or stool, and lighted by direct and reflected light. Moreover, the images coming from the endoscopic equipment itself can be interleaved, noisy, blurry and over- or under-exposed, and it can contain borders, sub-images and a lot of specular reflections (flares) caused by endoscope's light source. Therefore, detecting abnormalities and diseases in the GI tract is very different from detecting the objects from the surrounding world listed above. The GI tract can potentially be affected by a wide range of diseases with visible lesions (see Fig. 1d–e), but endoscopic findings may also include benign (normal) or man-made lesions. The most common diseases are gastric and colorectal cancer (CRC), which are both lethal when detected in a late stage. The 5-year



(a) Colon mucosa (normal) (b) Cecum (landmark) (c) Z-line (landmark)



(d) Polyps (disease) (e) Tumor (disease) (f) Blurry frames (notuseful)

Fig. 1 Example frames from human colon showing normal tissue (**a**)–(**c**), abnormal findings (**d**)–(**e**) and useless frames (**f**)

survival rate of CRC ranges from 93% in stage I to 8% in stage IV [29]. Consequently, early detection is crucial. There are several ways of detecting pathology in the GI tract, but systematic population-wide screening is important. However, current methods have limitations regarding sensitivity, specificity, access to qualified medical staff and overall cost.

In this scenario, both high precision and recall are important, but so is the frequently ignored system performance in order to provide feedback in real-time. The most recent and most complete related work is the Polyp-Alert polyp detection system [52], which can provide near real-time feedback during colonoscopies. However, it is limited to polyp detection, it uses edges, colors and texture in the images, and, at the moment, it is not fast enough for live examinations.

To further aid and scale such examinations, we have earlier presented EIR[1] [32, 37], an efficient and scalable automatic analysis and feedback system for medical videos and images. The system is designed to support endoscopists in the detection and interpretation of diseases in the GI tract. EIR has initially been tested in video analysis of the lower portions (large bowel) of the GI tract. However, our main objective is to automatically detect abnormalities in the whole GI tract. Therefore, we are developing a complete system for detection and in-frame position localization of different endoscopic findings like polyps, tumors, diseases and landmark objects (like the Z-line and cecum). The aim is to use next-generation-EIR for both (i) a computer assisted diagnosis tool for live endoscopic examinations and (ii) a future fully automated and scalable screening system used together with VCEs. These goals impose strict requirements on the accuracy of the detection to maximize number of true positives and to avoid false negatives (overlooking a disease), as well as low computational resource consumption to provide massive population screening with VCEs. The live-assisted system also introduces a real-time processing requirement defined

---

[1] In Scandinavian mythology, EIR is a goddess with medical skill.

as being able to process at least 30 HD frames per second, i.e., a common frame rate and resolution in modern endoscopic devices.

Our first version [32, 37] was developed for detection of polyps, i.e., possible cancer precursors, and it was built on content-based information retrieval methodology using global image features for image content analysis. In this paper, the next generation of our system is presented, where we extend our system using out-of-the-box and improved *deep learning* neural network approaches and multi-class global-feature classification methods for detection and localization of endoscopic findings. We evaluate our prototype by training new and improved classifiers that are based on various image-recognition approaches. We compare the performance of feature-based analysis and neural network-based analysis in terms of accuracy and real-time processing, and thereby evaluate the different approaches for feasibility of multi-class detection and colonic polyp localization in real use-case scenarios.

The results from our experimental evaluation show that, (i) the detection and localization accuracy can reach the same performance or outperform other current state-of-the-art methods, (ii) the processing performance enables frame rates for real-time analysis at high definition resolutions, (iii) the localization-system performance can be improved further using a combination of our basic localization algorithms and neural network approaches, (iv) in our experiments, the global-feature multi-class detection approach slightly outperforms the deep learning neural network approach both in training speed and detection performance, and (v) the system proves to be easily extended by adding new types of abnormalities. Thereby, a system based on global features seems to be preferable and gives better performance in multi-class object detection than given existing deep learning network approaches. For the localization, additional research is needed to achieve better performance using a combination of local feature detection and deep learning neural networks.

The rest of the paper is organized as follows: First, in Section 2, we briefly introduce our medical case study. Next, we present related work in the field and compare it to the presented system in Section 3. This is followed by a presentation of the complete system in Section 4. We present an evaluation of the system in Section 5, and in Section 6, we discuss two cases where our system will be used in two medical examinations by medical experts. Finally, we conclude our results in Section 7.

2 Gastrointestinal endoscopy

The GI tract can potentially be affected by various abnormalities and diseases. Some examples of possible findings are shown in Fig. 1b–e. CRC is a major health issue world-wide, and early detection of CRC or polyps as predecessors of CRC is crucial for survival. Several studies demonstrate that a population-wide screening program improves the prognosis and can even reduce the incidences of CRC [17]. As a consequence, in the current European Union guidelines, screening for colorectal cancer is recommended for all people over 50 years old [50]. Colonoscopy, a common medical examination and the gold standard for visualizing the mucosa and the lumen of the entire colon, may be used either as a primary screening tool or in a second step after other positive screening tests [25]. However, traditional rectal endoscopic procedures are invasive and may lead to great discomfort for patients, and extensive training of physicians and nurses is required to perform the examination. They are performed in real-time, and, therefore, it is challenging to scale the number of examinations to a large population. Additionally, the classical endoscopic procedures are expensive. In the US, for example, colonoscopy is the most expensive cancer screening process, with an annual cost of 10 billion dollars (1,100\$-6,000\$/person) [47], and a time consumption of about one medical doctor-hour and two nurse-hours per examination.

In our research, we aim for an algorithmic system that detects multiple mucosal pathologies in videos of the GI tract. The idea is to assist endoscopists (physicians, who are highly trained in the procedure) during live examinations. Additionally, alternatives to traditional endoscopic examinations have recently emerged with the development of non-invasive VCEs. The GI tract is visualized using a pill-sized camera (available from vendors such as Medtronics/Given and Olympus) that is swallowed and then records a video of the entire GI tract. The challenge in this context is that medical experts still need to view the full-length video. Our system should provide a scalable tool that can be used in a first-order population screening system where the VCE-recorded video is used to determine whether an *additional* traditional endoscopic examination is needed or not. As a first step, we target the detection and the localization of colorectal polyps, which are known precursors of CRC (see for example Fig. 1d). The reason for starting with this scenario is that most colon cancers arise from benign, adenomatous polyps (around 20%) containing dysplastic cells, which may progress to cancer. Detection and removal of polyps prevent the development of cancer, and the risk of getting CRC in the following 60 months after a colonoscopy depends largely on the endoscopist's ability to detect polyps [20]. Next, we extend our system to support detection of multiple abnormalities and diseases of the GI tract (see Fig. 1) by training the classifiers using multi-class datasets.

3 Related work

Detection of diseases in the GI tract has so far primarily focused on polyps. This is most probably due to the lack of alternative data in the medical field, but also that polyps are precursors of CRC. Several algorithms, methods and partial systems have, at first glance, achieved promising results [37] in their respective testing environment. However, none of the related works is able to perform real-time detection or support doctors by computer-aided diagnosis in real-time during colonoscopies. Furthermore, all of them are limited to a very specific use case, which in most cases is polyp detection for a specific type of camera [37]. Furthermore, in some cases, it is unclear how well the approach would perform as a real system used in hospitals. Most of the research conducted in this field uses rather small amounts of training and testing data, making it difficult to generalize the methods beyond the specific cleansed and prepared datasets and test scenarios. Therefore, overfitting for the specific datasets can be a problem and can lead to unreliable results.

The approach from Wang et al. [52] is the most recent and probably best-working system in the field of polyp detection. This system, called Polyp-Alert [52], is able to give near real-time feedback during colonoscopies. It uses an advanced edge-finding procedure to locate visual features and a rule-based classifier to detect an edge along the contour of a polyp. The system can recognize the same polyp across a sequence of video frames and can process up to 10 frames per second. The researchers report a performance of 97.7% correctly detected polyps with around 4.3% of frames incorrectly marked as containing polyps. Their results are based on a dataset that consists of 53 videos taken from different colonoscopes. Despite the promising polyp detection rate, the relatively high false positive rate makes the overall system detection performance not good enough for medical use cases. Unfortunately, the dataset used in this research is not publicly available, and therefore, a direct detection-performance comparison with our system is not possible. Moreover, most of the existing publications about polyp detection systems (see Tables 6 and 7 in Section 5) report detection accuracy on a per-polyp basis, counting the fact of successfully detected or missed polyp across the number of frames or even across the full video, which makes it difficult to perform a fair comparison. In our evaluation, we use a per-frame polyp detection and localization performance measurement. This gives a more realistic and better estimation of the performance of the developed method in the medical domain.

Other promising polyp detection approaches utilize quite old, but recently reborn neural networks and their advanced implementation called deep learning neural networks. Neural networks are conceptually easy to understand, and large amounts of research has been done in this direction in the last years. Results recently reported on, for example, the ImageNet dataset, look promising [13] in the areas of indexing, retrieving, organizing and annotating multimedia data. Despite the fact that the neural network model training process is very complicated and time-consuming [12], their ability to detect and localize various objects can potentially help us to improve our system. However, such an improvement is possible only after careful investigation, to ensure that our system will still run in real-time and be able to deal with the required amount of lesion categories. This is important since we deal with patient health, and the outcome can make the difference between life and death.

Most modern deep learning frameworks state that they can be used out-of-the-box for different types of input data. This statement sounds promising, but most state-of-the-art neural networks in multimedia research are designed to process images from everyday life, like cats, dogs, bicycles, cars, pedestrians, etc. It needs to be proven that they can be used in medical domains, because it is difficult to evaluate their performance and robustness properly [28] due to the lack of relevant training and test data. In fact, obtaining such datasets is one of the biggest challenges related to deep learning approaches in connection with the medical field, due to a lack of medical experts needed to annotate data, and legal and ethical issues. Some common conditions, like colon polyps, may already have the number of collected images and videos required to perform training of a neural network, while other endoscopic findings, like tattoos from previous endoscopic procedures (black-colored parts of the mucosa), are not that well documented, but still interesting to detect [40]. Recent research [8] on the topic of transfer learning promises a solution for the problem of insufficient amounts of available training data. Transferring the knowledge learned by the deep network on a large dataset, e.g. ImageNet, to train a specialized network on a small medically oriented dataset, together with a saliency prediction used to emphasize key image points, can result in better performance of the endoscopic finding detection and localization. Thus, in this research, we perform some preliminary experiments to see how neural networks can deal with small training datasets.

In summary, related work primarily targets specialized problems or elements of the more general, holistic medical problem we are attempting to solve. Existing systems are either (i) too narrow for a flexible, multi-disease detection system; (ii) have been tested on limited datasets too small to show whether the method would work in a real scenario, or; (iii) provide a processing performance too low for a real-time system or ignore the system performance entirely. Last, but not least, we are targeting a holistic end-to-end system where a VCE that traverses the entire tract with its video signals is algorithmically analyzed. To solve the fundamental systems problems, we are targeting and developing a close to fully automated, accurate, low false positive, scalable, privacy-preserving and low-cost screening system that will, if we may say so, have significant potential impact on the society.

4 The EIR system

Our objective is to develop a system that supports doctors in multi-disease detection in the GI tract. The system must (i) be easy to use and less invasive for the patients than existing methods, (ii) support multiple classes of detected GI objects, (iii) be easy to extend to new different diseases and findings, (iv) handle multimedia content in-real time (30 frames per second or more for Full HD videos), (v) be usable for real-time computer-aided diagnosis, (vi) achieve high classification performance with minimal false-negative classification results and (vii) have a low computational resource consumption. These properties potentially provide a scalable system with regard to reduced number of specialists required for a larger population, and dramatically increased number of users potentially willing to be screened. Therefore, EIR consists of three parts: the annotation subsystem [2], the detection and automatic analysis subsystem and the visualization and computer-aided diagnosis subsystem [35].

The subsystems for algorithmic analysis are designed in a modular way, so that they can be extended to different diseases or subcategories of diseases, as well as other tasks like size determination, etc. Currently, we have implemented two types of analysis subsystems: the detection subsystem that detects different irregularities in video frames and images, and the localization subsystem that localizes the exact position of the disease (only polyp localization is supported at the moment) in the frame. The detection subsystem is not designed to determine the location of the detected irregularity. The exact lesion position finding is done by the localization subsystem, so that we can use the same localization subsystem for different detection subsystems. The localization subsystem uses the output of the detection system as input and processes only frames marked as containing a localizable disease.

4.1 Detection subsystem

The detection subsystem performs lesion recognition and classification. It is intended for abnormality-presence detection without searching for the precise position of the lesion. The detection is performed using various visual similarity finding techniques. For each lesion that has to be detected, we use a set of reference frames that contains examples of this lesion occurring in different parts of the GI tract. This set can be seen as the model of the specific disease. We also use sets of frames containing examples of all kinds of healthy tissue, normal findings like stool, food, liquids, etc. The final goals of the detection subsystem is to decide if this particular frame analyzed contain any lesion or not, and to detect the exact type of the lesion. The detection system is designed in a modular way and can easily be extended with new diseases. This would, for example, allow not only to detect a polyp, but to distinguish between a polyp with low or high risk for developing CRC by using the *NICE* classification.[2]

4.1.1 Basic EIR system

In our previous work, we presented our basic EIR system [32, 36, 37] that implements a single-class global-feature-based detector able to recognize the abnormalities in a given video frame. Global image features were chosen, because they are easy and fast to calculate, and the exact lesion's position is not needed for detection, i.e., identifying frames that contain a disease. We showed that the global features we chose, Tamura feature [45] and Joint Composite Descriptor (JCD) [53], which is a combination of Fuzzy Color and Texture Histogram (FCTH) [10] and Color and Edge Directivity Descriptor (CEDD) [9], can indeed outperform or at least reach the same results as local features.

---

[2] http://www.wipo.int/classifications/nice/en/

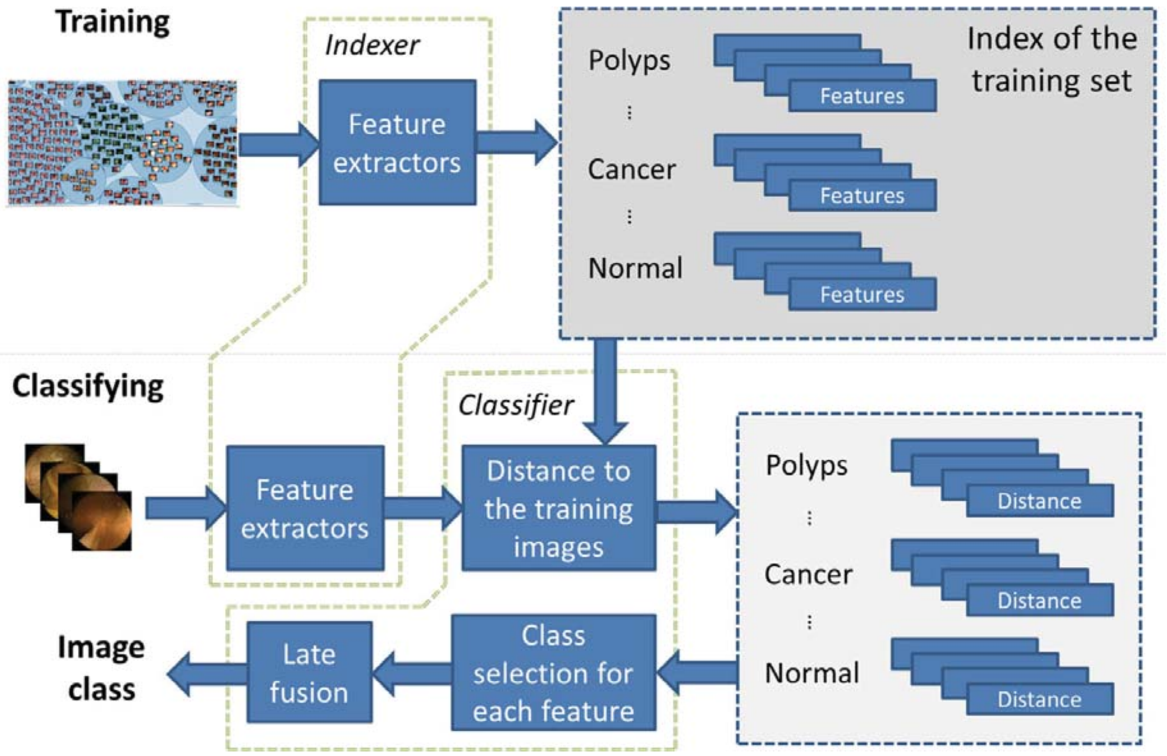


Fig. 2 Detailed steps for the multi-class global-feature-based detection implementation

The basic algorithm is based on an improved version of a search-based method for image classification. The overall structure and the data flow in the basic EIR system is depicted in Fig. 2. First, we create the index containing the visual features extracted from the training images and videos, which can be seen as a model of the diseases and normal tissue. The index also contains information about the presence and type of the disease in the particular frame. The resulting size of the index is determined by the feature vector sizes and the number of required training samples, which is rather low compared to other methods. Thus, the size of the index is relatively small compared to the size of the training data, and it can be easily fit into main memory on a modern computer. Next, during the classification stage, a classifier performs a search of the index for the frames that are visually most similar to a given input frame (see Section 4.1.3 for a detailed description of the method). The whole basic detector is implemented as two separate tools, an indexer and a classifier. We have released the indexer and the classifier as an open-source project called *OpenSea*[3] [37].

The indexer is implemented as a batch-processing tool. Creating the models for the classifier does not influence the real-time capability of the system and can be done off-line, because it is only done once when the training data is first inserted into the system. Visual features to calculate and store in the indexes are chosen based on the type of the disease because different sets of features or combinations of features are suitable for different types

---

[3] https://bitbucket.org/mpg_projects/opensea

of diseases. For example, bleeding is easier to detect using color features, whereas polyps require shape and texture information.

The classifier can be used to classify video frames from an input video into as many classes as the detection subsystem model consists of. The classifier uses indexes generated by the indexer. In contrast to other classifiers that are commonly used, this classifier is not trained in a separate learning step. Instead, the classifier searches previously generated indexes, which can be seen as the model, for similar visual features. The output is weighted based on the ranked list of the search results. Based on this, a decision is made. The classifier is parallelized and can utilize multiple CPU cores for the extraction of features and the searching in indexes. To increase performance even more, we implemented the most compute intensive parts of the system with GPU computation support.

4.1.2 Deep-EIR

The neural network version of EIR called Deep-EIR is based on a pre-trained convolutional neural network architecture and transfer learning [8]. We trained a model based on the Inception v3 architecture [43] using the ImageNet dataset [13] and then re-trained and fine-tuned the last layers. We did not perform complex data augmentation at this point and only relied on transfer learning. We are currently in the process of data collection, and for future work, we will also look into data augmentation and training a network from scratch using the newly collected data, which might lead to better results than transfer learning. Figure 3 gives a detailed overview of the complete pipeline for the neural network-based implementation of the detection.

Inception v3 achieves good results regarding single-frame classification and has reasonable computational resource consumption. The top one result error is 21.2%, and the top five error is 5.6% with less than 25 million parameters. The training of the Inception v3 network is performed from scratch using Google Tensorflow v1.2rc [1]. The training takes several weeks on a single modern computer with GPU support. Tensorflow is an open source framework that allows all kinds of numerical computations using graphs. Nodes within the flow graphs represent mathematical operations, and the edges represent data arrays (called tensors in Tensorflow). It is especially built to support scalable machine learning, which includes neural network-based architectures [1].

The trained Inception v3 model is then used in a retraining step. For this step, we follow the approach presented in [14]. Basically, we froze all the basic convolutional layers of

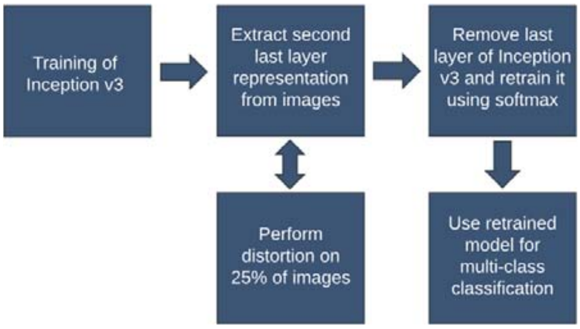

Training of Inception v3
Extract second last layer representation from images
Remove last layer of Inception v3 and retrain it using softmax
Perform distortion on 25% of images
Use retrained model for multi-class classification

Fig. 3 Detailed steps for the neural network approach based detection implementation

the network and only retrained the two top fully connected (FC) layers. The FC layers were retrained using the RMSprop [48] optimizer that allows an adaptive learning rate during the training process. After 1,000 epochs, we stopped the retraining of the FC layers and started fine-tuning the convolutional layers. For that step, we did the analysis of the Inception v3 model layer structure and decided to apply fine-tuning on the top two convolutional layers. This step finalizes the transfer-learning scenario and performs an additional tuning of all the NNs layers according to our dataset. For this training step, we used a stochastic gradient descent method with a low learning rate of 10^{-4} to achieve the best effect in terms of speed and accuracy [27]. This comes with the advantage that little training data is needed to train the network, which is an advantage for our medical use case. Additionally, it is fast, requiring just about one day to retrain the model. The re-trainer is based on an open source implementation of Tensorflow.[4] To increase the number of training samples, we also performed distortion operations on the images. Specifically, we performed random cropping, random rescaling and random change of brightness. The grade of distortion was set to 25% per image. After the model has been retrained, we use it for a multi-class classifier that provides the top five classes based on probability for each class.

4.1.3 Multi-class global-feature-based EIR

The new multi-class global-feature-based version of EIR is based on the initial version of EIR with some extensions. The basic search-based classification part of EIR is used to create a classifier for each disease that we want to classify. Figure 2 gives a detailed overview of the classifier's pipeline for the global-feature-based implementation of the detection. The difference to the basic EIR version is that the ranked lists of each search-based classifier are then used in an additional classification step to determine the final class.

For features extraction in the detection step and for the training procedure, the indexing is performed using the basic EIR indexer implementation [32, 37]. The same set of two global features, namely Tamura and JCD, is used. These features were selected by a simple features efficiency estimation by testing different combinations of features on smaller reference datasets to find the best combinations in terms of processing speed and classification accuracy. The selected features can be combined in two different ways. The first is called feature values fusion or early fusion, and it basically combines the feature value vectors of the different features into a single representation before they are used in a decision-making step. The second one is called decision fusion or late fusion where the features are combined after a decision-making step. Our multi-class global-feature-based approach implements feature combination using the late fusion.

During the detection step, a term-based query from the hashed feature values of the query image is created for each image, and a comparison with all images in the index is performed, resulting in a ranked list of similar images. The ranked list is sorted by a distance or dissimilarity function associated with the low-level features. This is done by computing the distance between the query image and all images in the index. The distance function for our ranking is the Tanimoto distance [46]. A smaller distance between an image in the index and the query image means a better rank [46]. The final ranked list is used in the classification

---

[4] https://github.com/eldor4do/Tensorflow-Examples/blob/master/retraining-example.py

step, which implements a simple k-nearest neighbors algorithm [4]. This algorithm can be used for supervised and unsupervised learning, two or multi-class classification and different types of input data ranging from features extracted from images to videos to meta-data. Its main advantages are its simplicity, that it achieves state-of-the-art classification results and that it is very fast in terms of processing time.

For the final classification, we use the random forest classifier [6], an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. A decision tree can be seen as a classifier, which basically performs decision-based classification on the given data. To get the final class, the classifier combines decision trees into a final decision implementing a late fusion for the multi-class classification. The advantage of the random forest algorithm is that the training of the classifier is very fast because the classification steps can be parallelized since each tree is processed separately. Additionally, it is shown that the random forest is very efficient for large datasets due to the ability to find distinctive classes in the dataset and also to detect the correlation between these classes. The disadvantage is that the training time increases linearly with the number of trees, which means a longer training time when many trees are used at the same time. However, this is not a problem for our use-case since the training is done offline, where time is less critical. Our implementation of the random forest classifier uses the version provided by the Weka machine learning library[5] [16], which is a collection of algorithms for machine learning and data mining. We chose the random forest approach, because it is fast and achieves good results [49]. It is important to point out that for this step, another classification algorithm can also be used.

4.2 Localization subsystem

The localization subsystem is intended for finding the exact positioning of a lesion, which is used to show markers on the frame containing the disease. This information is then used by the visualization subsystem. All images that we process during the localization step come from the positive frames list generated by the detection subsystem. Processing of the images is implemented as a sequence of intra-frame pre- and main-filters. Pre-filtering is needed because we use local image features to find the exact position of objects in the frames. Lesion objects or areas can have different shapes, textures, colors and orientations. They can be located anywhere in the frame and also partially be hidden and covered by biological substances, like seeds or stool, and lighted by direct light. Moreover, the image itself can be interlaced, noisy, blurry and over- or under-exposed, and it can contain borders and sub-images. Apart from that, images can have various resolutions depending on the type of endoscopy equipment used. Endoscopic images usually have a lot of flares and flashes caused by a light source located close to the camera. All these nuances affect the local feature-based detection methods negatively and have to be specially treated to reduce localization precision impact. In our case, several sequentially applied filters are used to prepare raw input images for the following analysis. These filters are border and sub-image removal, flare masking and low-pass filtering. After pre-filtering, the images are ready to be used for further analysis.

---

[5] http://www.cs.waikato.ac.nz/ml/weka/

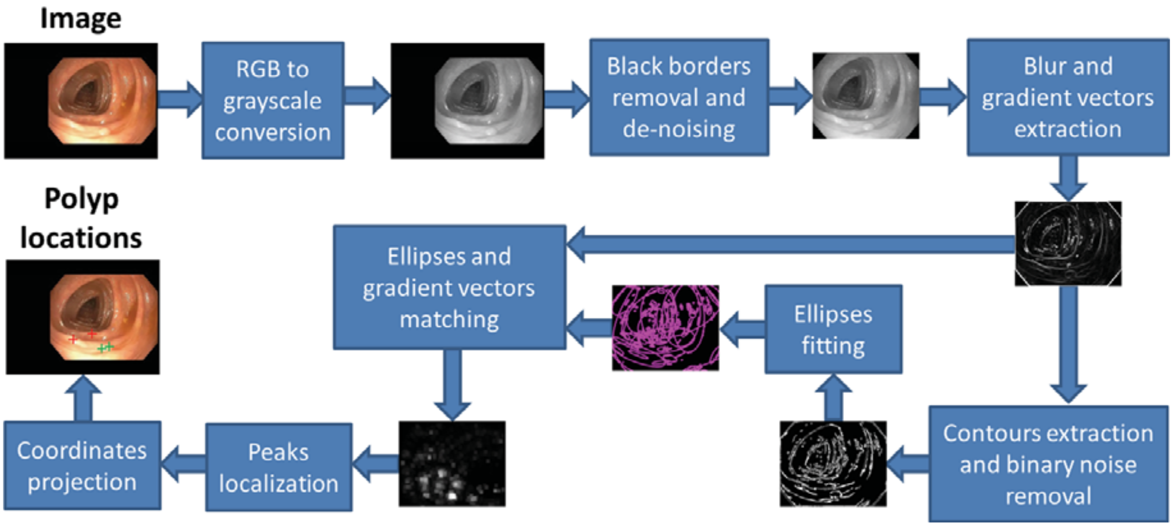


Image
RGB to grayscale conversion
Black borders removal and de-noising
Blur and gradient vectors extraction
Polyp locations
Ellipses and gradient vectors matching
Ellipses fitting
Contours extraction and binary noise removal
Coordinates projection
Peaks localization

Fig. 4 Detailed steps of the basic EIR localization algorithm implementation

4.2.1 *Basic EIR system*

Previously, we have implemented the localization of colon polyps using our hand-crafted approach based on local image features [35]. The main idea of the localization algorithm is to use the polyp's physical shape to find the exact position in the frame. In most cases, the polyps have the shape of a hill located on a relatively flat underlying surface or the shape of a more or less round rock connected to an underlying surface with stalks of varying thickness. These polyps can be approximated with an elliptically shaped region consisting of local features that differ from the surrounding tissue with high probability. To detect these types of objects, we process the frames marked by the detection subsystem as containing polyps by a sequence of various image processing procedures, resulting in a set of possible abnormality coordinates within each frame. Figure 4 gives a detailed overview of a localization pipeline for the basic EIR algorithm implementation. The pipeline consists of the following steps: non-local means de-noising [7]; 2D Gaussian blur and 2D image gradient vector extraction; border extraction by gradient vector threshold binarization; border line isolated binary noise removal; estimation of ellipses locations; ellipse size estimation by analyzing border pixel distribution; ellipse fitting to extracted border pixels; selection of a predefined number of non-overlapping local peaks and outputting their coordinates as possible polyp locations. For the possible locations of ellipses, we use the coordinates of local maxima in the insensitivity image, created by additive drawing of straight lines starting at each border pixel in the direction of its gradient vector. Ellipse fitting is then performed using an ellipse fitting function [15]. This version of the subsystem is implemented in C++, and it uses the OpenCV[6] open source library for routine image content manipulation and the CUDA[7] toolkit for GPU computation support.

4.2.2 *Deep-EIR*

The existing localization scheme can be extended to support different diseases by implementation of lesion-specific shape, color and texture detection, but such an extension

---

[6] http://opencv.org/

[7] http://developer.nvidia.com/cuda-toolkit

requires experimental studies for each new type of abnormality. In order to reduce the system improvement costs, we performed an evaluation of two universal object localization frameworks, based on deep learning neural network approaches. First is TensorBox[8] [41], which extends Google's reference implementation of the machine-learning framework called Tensorflow [1]. Second approach is based on the Darknet [33] open-source deep learning neural network implementation called YOLO[9] [34]. Both of these frameworks are designed to provide not only object detection, but also object localization inside frames. They implement GPU-accelerated deep learning algorithms that can work with near to real-time performance and provide the capability of locating various objects out-of-the-box.

The TensorBox approach introduces an end-to-end algorithm for detecting objects in images. As input, it accepts images and directly generates a set of object bounding boxes as output. The main advantage of the algorithm is the capability of avoiding multiple detections of the same object by using a recurrent neural network (RNN) with long short-term memory (LSTM) units together with fine-tuned image features from the implementation of a convolutional neural network (CNN) for visual objects classification and detection called GoogLeNet [42].

The Darknet-YOLO approach introduces a custom CNN, designed to simultaneously predict multiple bounding boxes and class probabilities for these boxes within each input frame. The main advantage of the algorithm is that the CNN sees the entire image during the training process, so it implicitly encodes contextual information about classes as well as their appearance, resulting in a better generalization of objects' representation. The custom CNN in this approach is also inspired by the GoogLeNet [42] model.

As initial models for both approaches, we used database models pre-trained on ImageNet [19] . Our custom training and testing data for the algorithms consists of frames and corresponding text files describing ground truth data with defined rectangular areas around objects: a JSON file for TensorBox and one text file per frame for Darknet-YOLO. Ground truth data was generated using a binary-masked frame set (example shown in Fig. 5) by the localization validation software used in our experimental studies. Both frameworks were trained using the same training dataset, where all frames contained one or more visible polyps. No special filtering or data preprocessing was used, thus the training dataset contained high quality and clearly visible polyp areas as well as blurry, noisy, over-exposed frames and partially visible polyps. The models were trained from scratch using corresponding default-model training settings [34, 41]. After the training, the test dataset was processed by both neural networks in testing mode. As a result, the frameworks output JSON (TensorBox) and plain-text (Darknet-YOLO) files containing sets of rectangles, one set per frame, marking possible polyp locations with corresponding location confidence values. These results have been processed using our localization algorithms.

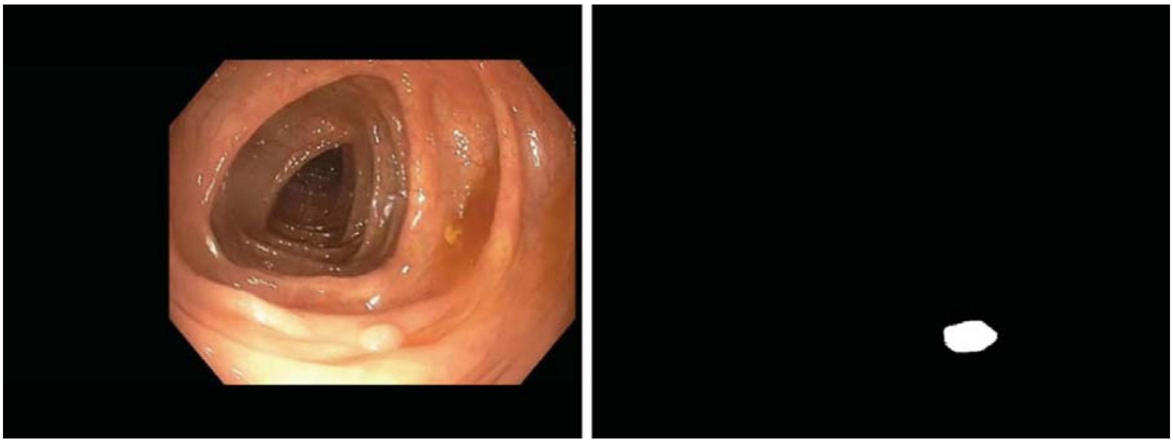
4.3 Visualization and computer aided diagnosis subsystem

The visualization subsystem is developed as a flexible multi-purpose tool. First, it should help in evaluating the performance of the system and get insights into why things work well or not. Second, it can be used as a computer-aided diagnostic system for medical experts. Third, it should help us in the creation of new datasets, allow us to extend the number of detected diseases and help doctors to create annotations in a time-saving manner. Previously,

---

[8] https://github.com/Russell91/TensorBox

[9] https://github.com/pjreddie/darknet

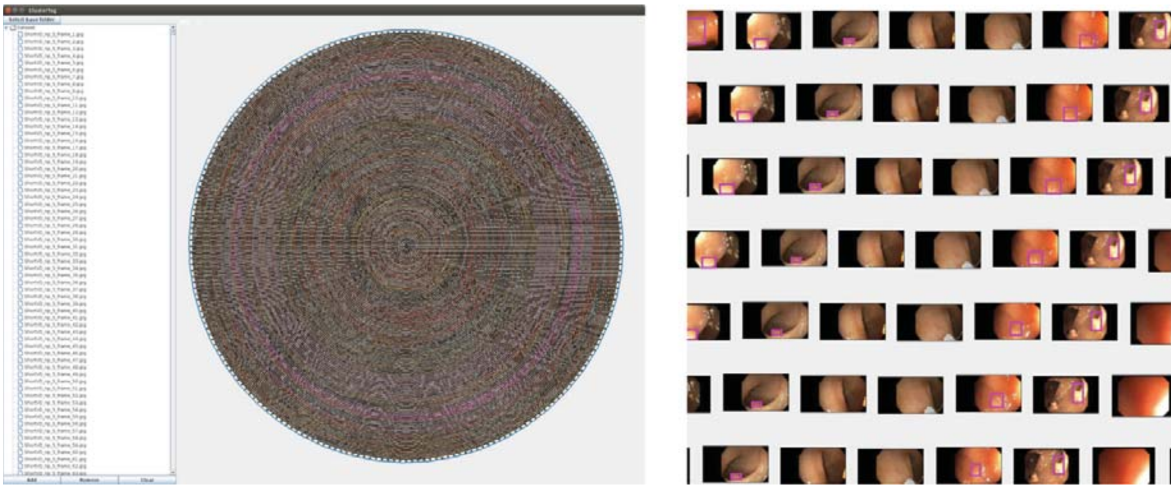


(a) Frame with polyp (b) Polyp ground truth

Fig. 5 Example frames showing polyp and it’s body ground truth area. This is an example of polyps localization task complexity. Polyp body has the same color, texture properties and light flares as surrounding normal mucosa

we have developed the TagAndTrack subsystem [2] that can be used for visualization and computer-aided diagnosis. We developed a web-based visualization toolkit that can be used to support medical experts while being very easy to use and distribute. This tool takes the output of the detection and localization subsystems and creates a web-based representation of the detection and localization results. The web-based visualization is then combined with a video sharing and annotating platform where doctors are able to watch, archive, annotate and share information. To break through low availability of high quality training and testing datasets for different GI track diseases, we developed a new ClusterTag application for the visualization subsystem. The main purpose of ClusterTag is to provide an easy-to-use and convenient user interface to huge image and video frame collections captured during endoscopic procedures, including conventional colonoscopies and VCEs.

Figure 6 illustrates our ClusterTag application while processing a dataset containing 36, 476 images with the exact lesion areas marked. The application implements image and



(a) Main window. (b) Zooming closer to images set.

Fig. 6 ClusterTag application usage example. The loaded dataset contains 36, 476 images with ground truth (marked by pink rectangles on images)

ground truth loading and analyzing, image tagging, creation and editing of ground truth data, global feature extraction and semi-automatic dataset clustering using our previously developed algorithms [38]. With the main focus on the interactive visual representation of huge image collections, the visualization module helps users create and interact with the new or already defined clusters. We use the Weka library to help the user in building clusters. For the image attribute extraction required for machine-learning-based classification we use global image features, which are extracted using the image retrieval framework called LIRE.[10] In our approach, we use global features describing the image in terms of different visual attributes, such as sharpness, color distribution and histogram of brightness. A detailed description of the used global features, the corresponding clustering algorithm and the clustering performance metrics can be found in [38]. Both the WEKA and LIRE libraries can be easily replaced by other machine learning or feature extraction libraries if desired.

Applying unsupervised clustering on huge unsorted and unannotated datasets significantly reduces the amount of work required from skilled doctors during image labeling and grouping. Together with unsupervised clustering, our application provides the users with the ability of tagging and analyzing multiple single images at once and putting them into appropriate groups together. The ClusterTag application is released as open-source software[11] and might help other research groups in the creation and analysis of new datasets.

5 Evaluation

For our experimental evaluation, we use two different use-cases. First, we evaluated the performance of our multi-class classification and detection algorithms in automated colonoscopy video processing. Here, we tested our system using six different classes of endoscopic findings that can be found in the colon (shown in Fig. 1). The classes to be detected are (a) frames with normal colon mucosa (healthy colon wall), (b) frames of the cecum area which is an intraperitoneal pouch that is considered to be the beginning of the colon (an anatomic landmark helping doctors and VCE video analysis algorithms to orientate in the colon), (c) frames displaying the Z-line which is the gastroesophageal junction that joins the esophagus to the stomach (an anatomic landmark), (d) frames containing one or more polyps, (e) frames with visible tumor areas, and (f) useless blurry frames without any visible and recognizable objects. Thus, the developed multi-class classification and detection system should split all the video frames into six classes that can be observed in the human GI tract. The developed method allows us to implement a new generation of endoscopy video processing systems able to efficiently detect various lesions of the GI tract.

Second, we evaluated the performance of the state-of-the-art object localization approaches based on deep learning algorithms, and then we compared it with our basic polyp localization algorithm. In this use-case, we compared the ability of different methods to find the location of polyps inside a frame. The main goal of this evaluation is to decide if we can improve the polyp localization performance of our system using a combination of different algorithms.

During the evaluation, wherever it was possible, we compared the performance of our method with the best state-of-the-art competitors. Nevertheless, a direct comparison is hard

---

[10] http://www.lire-project.net/

[11] https://bitbucket.org/mpg_projects/clustertag

as different datasets and detection measures are used in state-of-the-art system evaluations. Thus, we compared the metrics we found in the relevant publications.

For all of the subsequent measurements, we used the same computer. It has an Intel Core i7-6700K CPU running at 4.00GHz, 16 GB of RAM, a GeForce GTX TITAN X GPU, and it runs a 64-bit Ubuntu Linux v16.04.

5.1 Multi-class classification

In the multi-class classification experiments, we used cross-validation because of the relatively small number of images in the annotated dataset. For the performance measurement, we used the standard tool from WEKA for evaluating multi-class classifiers. This tool uses the ground truth to compute a confusion matrix and the common standard metrics: recall (sensitivity), precision, specificity, accuracy and F1 score. We created a new dataset from colonoscopy images that we got from Vestre Viken Hospital, Norway. From the whole unannotated dataset, we manually selected 50 different frames of 6 different classes (described in Section 2): blurry frames, cecum, normal colon mucosa, polyps, tumor, and Z-line. The selected frames were used to create 10 separate datasets, each containing training and test subsets with equal numbers of images. Training and test subsets were created by equally splitting random-ordered frame sets for each of the 6 classes. The total number of frames used in this evaluation is 300: 150 in the training subsets and 150 in the test subsets. Each training and test subset contains 25 images per class. Multi-class classification is then performed on all 10 splits and then combined and averaged. Following this strategy, an accurate enough estimation about the performance can be made even with a smaller number of images.

5.1.1 Deep-EIR

First, we performed an evaluation of Deep-EIR that implements the deep learning neural network multi-class detection approach. Table 1 shows the resulting confusion matrix. The detailed performance metrics presented in Table 2 and the results can be considered as good, they confirm that Deep-EIR performs well. All blurry and Z-line frames were classified correctly. Cecum and normal colon mucosa were often cross-mis-classified, which is a normal behavior, because from a medical point of view, normal colon mucosa is part of the cecum, and under real-world circumstances, this would not be a relevant mistake. Interesting polyps

Table 1 A confusion matrix for the six-classes detection performance evaluation for the Deep-EIR detection subsystem

		Detected class					
		Blurry	Cecum	Normal	Polyps	Tumor	Z-line
Actual class	Blurry	**250**	0	0	0	0	0
	Cecum	0	**183**	64	3	0	0
	Normal	0	34	**197**	19	0	0
	Polyps	1	17	45	**183**	4	0
	Tumor	0	0	1	4	**245**	0
	Z-line	0	0	0	0	0	**250**

Bold numbers shows the correct detection result for each class

Table 2 Performance evaluation of the six-classes detection for the Deep-EIR detection subsystem

	True Pos.	True Neg.	False Pos.	False Neg.	Recall (Sensitivity)	Precision	Specificity	Accuracy	F1 score
Blurry	250	1249	1	0	100.0%	99.6%	99.9%	99.9%	**99.8%**
Cecum	183	1199	51	67	73.2%	78.2%	95.9%	92.1%	**75.6%**
Normal	197	1140	110	53	78.8%	64.2%	91.2%	89.1%	**70.7%**
Polyps	183	1224	26	67	73.2%	87.6%	97.9%	93.8%	**79.7%**
Tumor	245	1246	4	5	98.0%	98.4%	99.7%	99.4%	**98.2%**
Z-line	250	1250	0	0	100.0%	100.0%	100.0%	100.0%	**100.0%**
Overall	1308	7308	192	192	87.2%	87.2%	97.4%	95.7%	**87.2%**

Bold numbers shows the balanced F-score of each proposed method

and tumors were detected correctly in most cases, as well as the Z-line landmark, which is important for our medical use case.

5.1.2 Multi-class global-feature-based EIR

Second, we performed an evaluation of the multi-class global-feature-based EIR, which implements a global-feature multi-class detection approach. The multi-class global-feature-based EIR classifier allows us to use a number of different global image features for the classification. The more image features we use, the more precise the classification becomes. We generated indexes containing all possible image features for all frames of all different classes of findings from our training and test dataset. These indexes were used for multi-class classification, different performance measurements and also for leave-one-out cross-validation. Using our detection system, the built-in metrics functionality can provide information on the different performance metrics for benchmarking. Further, it provides us with the late fusion of all the selected image features and performs the selection of the exact class for each frame in test dataset. All used features are described in detail in [24].

Table 3 shows the resulting confusion matrix, which shows, like the Deep-EIR results, that the global feature-based detection approach performs well, too. Again, all blurry and Z-line frames were classified correctly. Cecum and normal colon mucosa were sometimes

Table 3 A confusion matrix for the six-classes detection performance evaluation for the multi-class global-feature-based EIR detection subsystem

		Detected class					
		Blurry	Cecum	Normal	Polyps	Tumor	Z-line
Actual class	Blurry	**250**	0	0	0	0	0
	Cecum	0	**226**	21	3	0	0
	Normal	0	85	**165**	0	0	0
	Polyps	0	10	8	**226**	6	0
	Tumor	0	0	0	8	**242**	0
	Z-line	0	0	0	0	0	**250**

Bold numbers shows the correct detection result for each class

Table 4 Performance evaluation of the six classes detection for the multi-class global-feature-based EIR detection subsystem

	True Pos.	True Neg.	False Pos.	False Neg.	Recall (Sensitivity)	Precision	Specificity	Accuracy	F1 score
Blurry	250	1250	0	0	100.0%	100.0%	100.0%	100.0%	**100.0%**
Cecum	226	1155	95	24	90.4%	70.4%	92.4%	92.1%	**79.2%**
Normal	165	1221	29	85	66.0%	85.1%	97.7%	92.4%	**74.3%**
Polyps	226	1239	11	24	90.4%	95.4%	99.1%	97.7%	**92.8%**
Tumor	242	1244	6	8	96.8%	97.6%	99.5%	99.1%	**97.2%**
Z-line	250	1250	0	0	100.0%	100.0%	100.0%	100.0%	**100.0%**
Overall	1359	7359	141	141	90.6%	90.6%	98.1%	96.9%	**90.6%**

Bold numbers shows the balanced F-score of each proposed method

cross-misclassified. Polyps and tumors were detected correctly in most cases. The detailed performance metrics are presented in Table 4 and can also be considered as good.

5.1.3 Deep-EIR vs multi-class global-feature-based EIR

The comparison of these two approaches shows that both approaches have equal excellent overall F1 score of 100% in Z-line detection. The global-feature approach with the 100% F1 score outperforms the neural network approach by a small margin in blurry frame detection. The neural network F1 score detection for tumors is 98.2%, which is 1% better than the global-feature approach. Detection of other classes is better for the global-feature approach, giving the F1 scores of 79.2% and 74.3% for cecum and normal mucosa. Most importantly for our case study, polyp detection performed much better using the global-feature approach, giving the 92.8% F1 score (13.1% better than the neural network approach).

The performance evaluation of the cross-validation for both multi-class classification approaches (see Table 5) confirms the high stability of the models used for the classification.

The processing performance of both Deep-EIR and global-feature-based EIR in terms of processing speed meets real-time demands with a good margin for the real-time medical use case. Both can process Full HD images at a frame rate of 30 frames per second.

Our experimental comparison of the Deep-EIR and the global-feature-based EIR of the detection system shows clearly that the global-feature approach outperforms the deep learning neural network approach and gives better accuracy for almost all target detection classes (except several cases of misclassification of tumors) in conjunction with high 92.8% and 97.2% F1 scores for the most important findings: polyps and tumors. Moreover, when a

Table 5 Performance evaluation of the cross-validation for the Deep-EIR and the multi-class global-feature-based EIR detection subsystems

Approach	Mean absolute error	Root mean squared error	Relative absolute error, %	Root relative squared error, %
Deep-EIR	0.07284	0.20574	26.21936	55.21434
Multi-class global-feature-based EIR	0.09242	0.19644	33.2672	52.7148

sufficiently large training dataset covering all possible detectable lesions of the GI tract is used, the proposed global-feature approach for multi-class detection requires relatively little time for training [35] compared to days and weeks for the deep learning neural network approach.

A comparison of Deep-EIR and global-feature-based EIR with existing competitive approaches is shown in Table 6. The basic-, Deep- and multi-class global feature-based EIR detector versions are depicted in the last table's rows. As one can see, the global feature-based EIR approach gives the best performance in terms of precision (90.6%), specificity (98.1%) and accuracy (96.9%), and comparable recall/sensitivity (90.6%). In other words, the results indicate that we can detect different classes of GI tract findings with a precision of almost 91%. If we compare this to the best performing system in Table 6, we see that Polyp-Alert reaches slightly higher detection accuracy on a different dataset. However, our system is faster and can detect colonoscopic findings in real-time, and furthermore, it is not designed and restricted to detect only polyps, it can detect multiple classes of diseases, and EIR can further be expanded to any additional diseases if we have the correct training data.

The performance comparison of different multi-class detection and classification approaches in terms of frame processing speed is depicted in Fig. 7. Deep-EIR, multi-class global feature-based EIR and basic EIR perform better in terms of speed than competitors. The single-class basic EIR detector can process up to 300 Full HD frames per second (for a GPU-accelerated implementation) [35]. Deep- and global feature-based EIR classifiers showed 30 frames per second, which fits our medical use case. For further processing speed improvements, we plan to implement additional GPU acceleration for a random-trees

Table 6 A performance comparison of GI findings detection approaches

Publ./System	Detection Type	Recall (Sensitivity)	Precision	Specificity	Accuracy	FPS	Dataset Size, images
Wang et al. [52]	polyp / edge, texture	97.70%	–	–	95.70%	10	1.8m
Wang et al. [51]	polyp / shape, color, texture	81.4%	–	–	–	0.14	1,513
Mamonov et al. [26]	polyp / shape	47%	–	90%	–	–	18,738
Hwang et al. [18]	polyp / shape	96%	83%	–	–	15	8,621
Li and Meng [23]	tumor / textural pattern	88.6%	–	96.2%	92.4%	–	–
Zhou et al. [54]	polyp / intensity	75%	–	95.92%	90.77%	–	–
Alexandre et al. [3]	polyp / color pattern	93.69%	–	76.89%	–	–	35
Kang et al. [21]	polyp / shape, color	–	–	–	–	1	–
Cheng et al. [11]	polyp / texture, color	86.2%	–	–	–	0.076	74
Ameling et al. [5]	polyp / texture	95%	–	–	–	–	1,736
Basic EIR [35]	polyps / 30 features	98.50%	93.88%	72.49%	87.70%	300	18,781
Deep-EIR	abnormalities / neural network	87.20%	87.20%	97.40%	97.50%	30	300
Multi-class global-feature-based EIR	abnormalities / 30 features	90.60%	90.60%	98.10%	96.90%	30	300

Not all performance measurements are available for all methods, but including all available information gives an idea about each method's performance

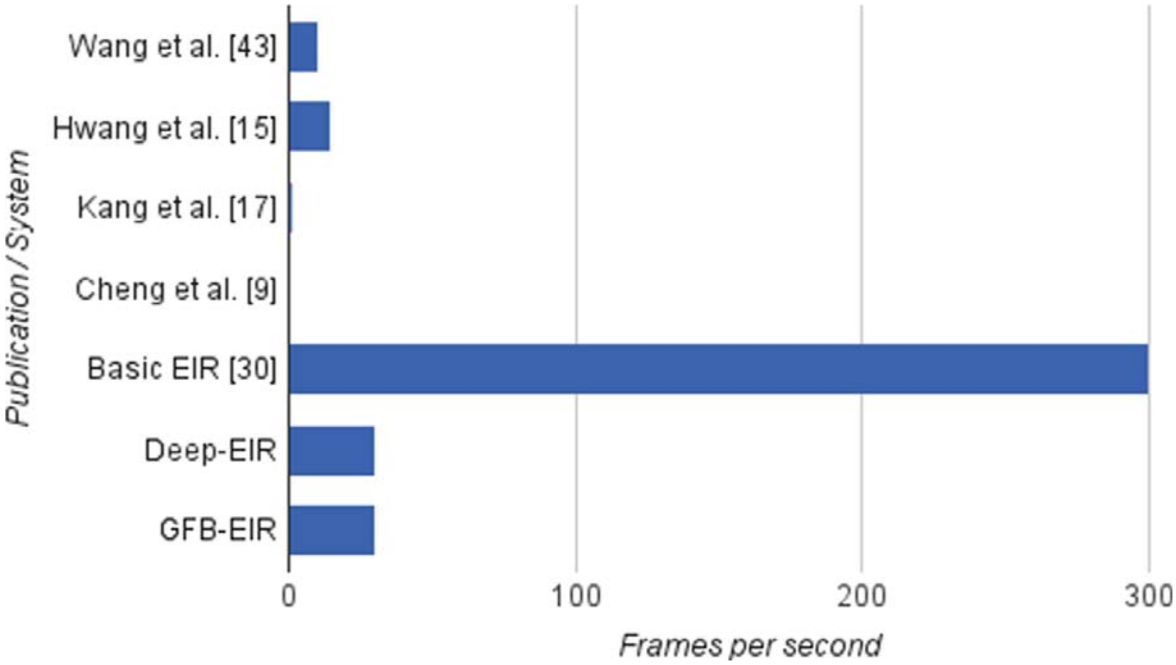


Fig. 7 The chart shows a comparison of different GI tract finding detection approaches. The presented Deep-EIR and multi-class global-feature-based EIR (GFB-EIR) systems show performance of 30 frames per second, which is higher comparing to other systems

classifier and feature index search, as we have for our initial polyp detection version of EIR [32].

5.2 Polyp localization

The multi-class dataset from Vestre Viken Hospital does not contain the ground truth for the localization of the findings. Therefore, in this experiment, we used the available ASU-Mayo Clinic polyp database.[12] It consists of training and test sets of images and videos with corresponding ground truth showing the exact polyp location areas. This was the biggest publicly available dataset (until recently, when the owners decided to withdrawn it from the public), consisting of 20 videos from standard colonoscopies with a total of 18,781 frames and different resolutions up to full HD [44]. For this particular evaluation, we selected only frames containing polyps, which gave us 8,169 frames in total: 3,856 in the training subset and 4,313 in the test subset. The frames with polyps contain various polyp types, fully visible and particularly hidden, clearly visible and blurry, clean and covered by stool. Figure 8 depicts variations in polyp sizes (in terms of number of pixels showing polyp bodies within images) across the datasets. As one can see, there are huge variations in polyp sizes in terms of video-frame pixels from very small up to one third of the full video frame size. This reflects real colonoscopy video-capturing scenarios and introduces a big challenge for object localization algorithms.

For the localization-performance measurement, we used the common metrics: recall (sensitivity), precision, specificity, accuracy and F1 score. To count the corresponding localization events correctly, we took into account that polyps can have different shapes, they are often not located in compact pixel space areas (in contrast to, e.g., people faces). The

---

[12] http://polyp.grand-challenge.org/

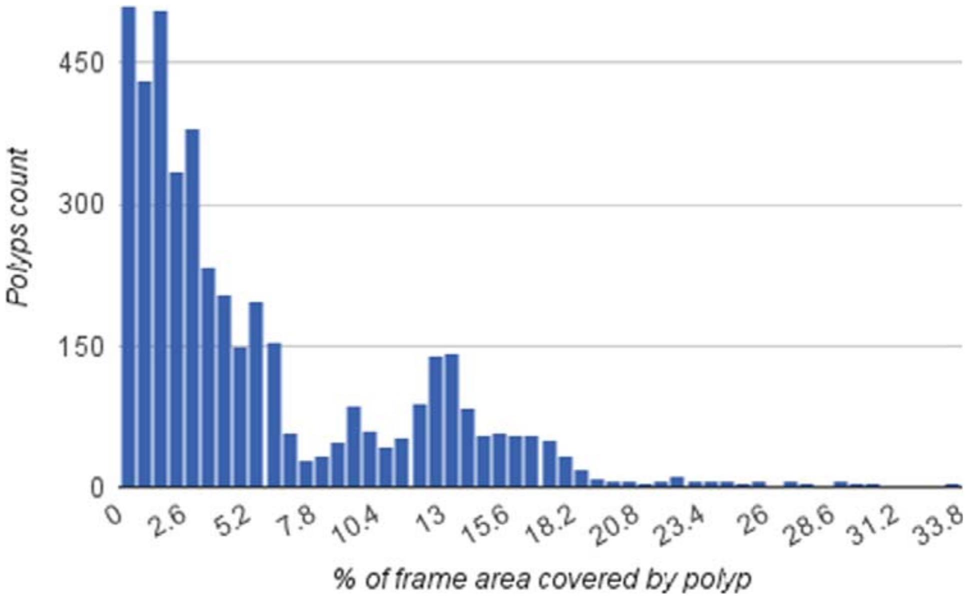


Fig. 8 The histogram shows huge variations in number of frame pixels, covered by polyp bodies, from very small up to one-third of full frame size across whole ASU-Mayo Clinic polyp database

shape of the polyps is marked in the ground truth data by binary masks. Before computing the localization subsystem performance, we need to figure out how to convert output of different localization algorithms into performance metrics. Our initial assumption (from practical experience) was to count each of the neural networks' location rectangles as a true positive localization event if and only if it covers at least 10% of the corresponding ground truth area. Otherwise, we count it as a false positive. In our use case, multiple detection of the same polyp does not improve medical outcome. Therefore, we count multiple true positives on the same polyp ground truth area as one true positive. Polyp misses are counted if, after processing all resulting rectangles for a particular frame, we still have one or more ground truth areas without corresponding true positives. We count such misses as false negatives. Thus, there is a possibility of multiple false negatives per one frame, in case we have multiple lesions in the same frame. In this experiment, we process only frames that contain one or more polyps. This means that we do not have true negatives. Therefore, specificity of the algorithms can be assumed as 100%. To check our assumptions about minimal coverage areas, we performed an initial performance evaluation and built a graph showing unfiltered output from neural networks. In our EIR system, the base localization algorithm outputs points instead of rectangular areas. Thus, we count a true positive if a point is located inside of a polyp ground truth area, keeping other rules the same. An example of a polyp localization algorithm output is depicted in Fig. 9. The polyp-location ground truth marked by light green ellipses is computed based on the ground truth binary masks (see Fig. 5) using the closest elliptical region approximation. Due to the limitations of the current version of the localization algorithm, it produces four possible polyp locations per frame without any location ranking. In this evaluation, we consider all four points as equal and always use all of them for calculating the performance metrics. These points are marked by the green and red crosses. The green crosses correspond to the true positive events, and the red crosses show the false positive events.

The deep learning neural network frameworks tested in this experiment require training before they are able to perform polyp localization. Thus, both networks were trained using their default model training parameters. For TensorBox, the neural network model training

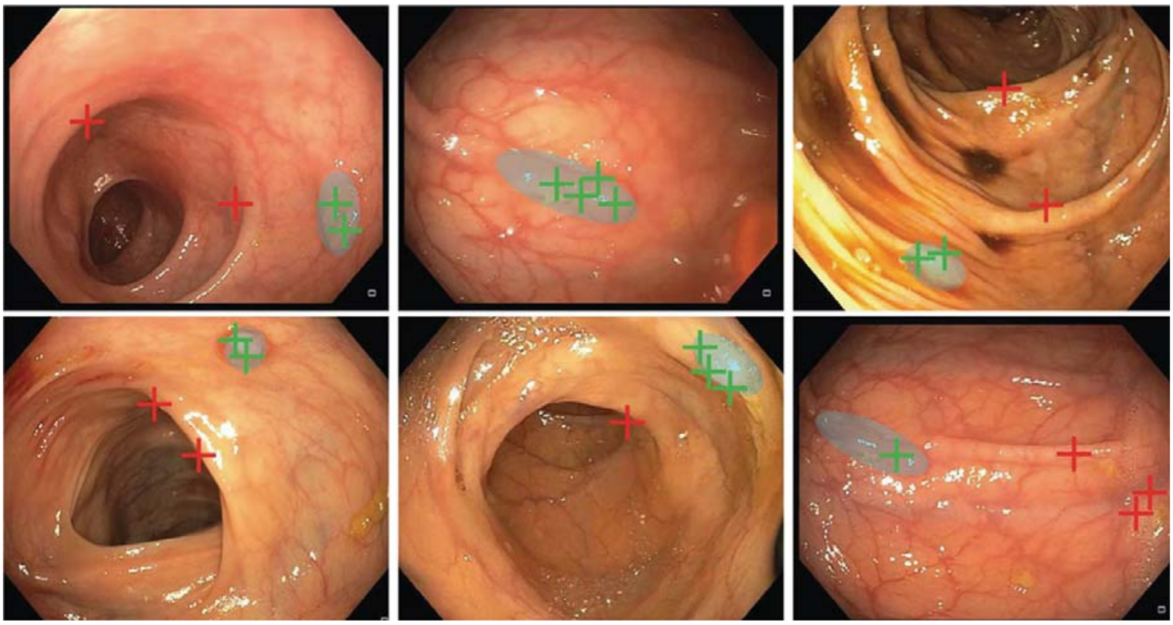


Fig. 9 Example of the polyp localization algorithm output. The current version of the algorithm produces four possible polyp locations per frame. The polyp location ground truth is marked by *light green* ellipses. The *green* crosses correspond to the true positives, the *red* correspond to the false positives

took 6.5 days, and for Darknet-YOLO, we needed 5.1 days. After the training, we performed model validation using the corresponding frameworks' routines, and the training dataset as input. The validation confirmed the correctness of the trained models for both TensorBox and Darknet-YOLO. The deep learning approaches are capable of correctly localizing polyps that were previously detected by the detection subsystem within the training dataset with 98% accuracy for the TensorBox model and 95% accuracy for the Darknet-YOLO model.

Next, we performed a main localization run of both frameworks on the test dataset and validation using the corresponding ground truth. Both TensorBox and Darknet-YOLO can be finely tuned by setting confidence threshold values, which limits the number of returned location rectangles to only highly confident ones. In order to investigate how the output of both can be affected by a confidence threshold value, it was set to zero during the first test run, which should give us the full unfiltered localization output. The reason for studying this dependency is that it is the only network tuning parameter in the unseen data process mode, which can help us to maximize their localization accuracy. Figure 10 shows a histogram of true polyps' area coverage by location boxes found by TensorBox. We counted only location boxes that cover at least one pixel of a true polyp area. As one can see, the histogram has clearly visible maximum around 16% coverage rate, followed by an exponential decrease to almost constant level. A comparable analysis with the same type of histogram for the Darknet-YOLO output is depicted in Fig. 11. We observe a similar distribution for coverage rate (higher than 10%). A much higher number of location rectangles with zero coverage rate indicates that TensorBox implements additional localization result filtering. Thus, the effect of the confidence threshold level adjustment cannot be as significant as for Darknet-YOLO, which has the expected output with a high number of location boxes covering small parts of true polyp areas. Therefore, Darknet-YOLO should show a strong response to confidence threshold level. For the following validation and performance evaluation of both frameworks, we used 10% as the threshold value for the minimal required polyp ground truth coverage for true positive events, i.e., 10% must be covered for the event

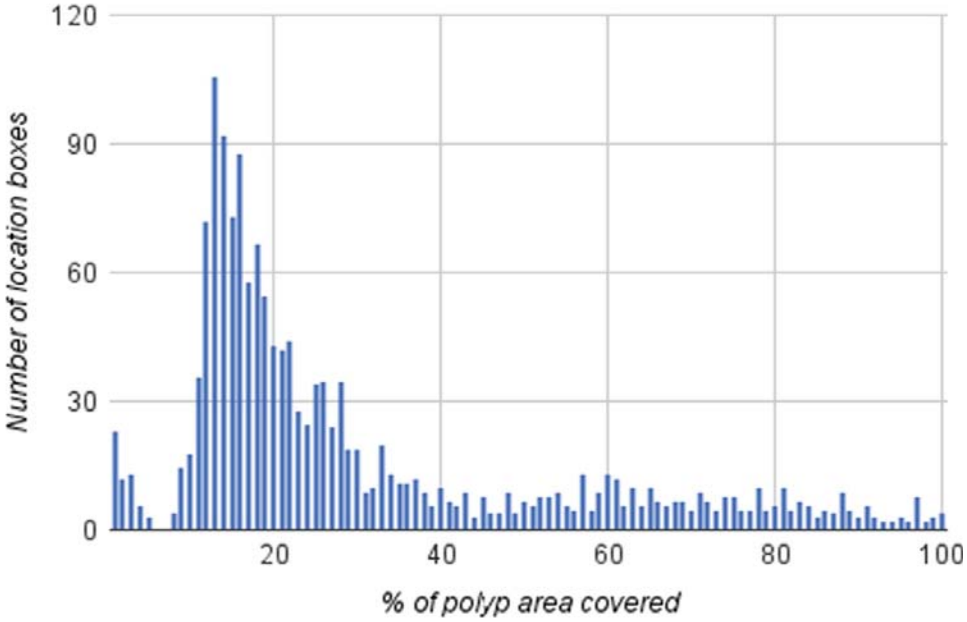


Fig. 10 The histogram shows polyps area coverage by location boxes found by the TensorBox localization algorithm with the maximum around 16% coverage rate with following exponential decrease to the almost constant level. The low number of found location rectangles around zero coverage rate is an evidence of some output results pre-filtering

to be counted. Figures 12 and 13 confirm our assumption about output result filtering in TensorBox. Its output contains a relatively small number of found locations with high number of highly-confident locations compared to Darknet-YOLO, which has a large number of low-confident locations, exactly as expected with the choice of a zero-confidence threshold.

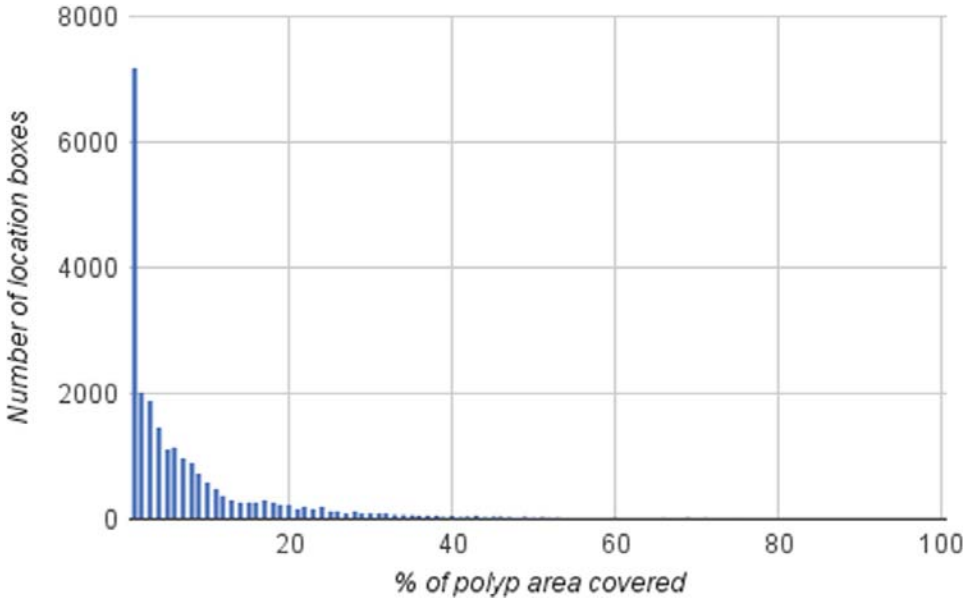


Fig. 11 The histogram shows polyps area coverage by location boxes found by the Darknet-YOLO localization algorithm with near to exponential distribution for coverage rate higher than 10%. The higher number of found location rectangles around zero coverage rate gives clear indications that algorithm output unfiltered results

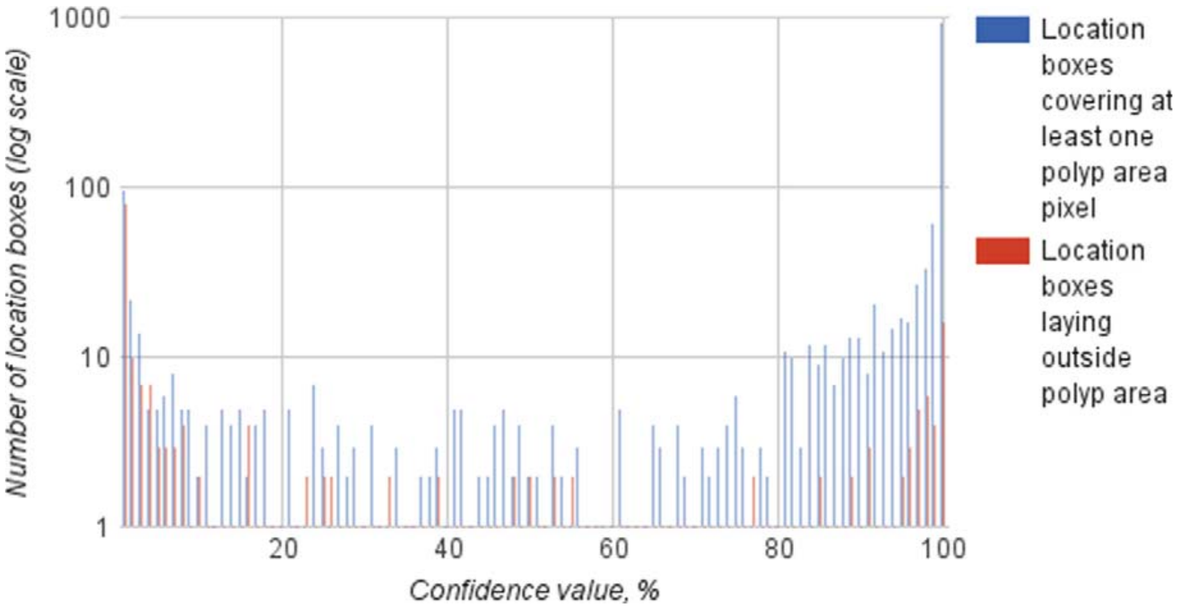


Fig. 12 The histogram shows confidence values for location boxes found by the TensorBox localization algorithm. It shows the relatively low number of found locations with high number of highly-confident locations

The performance results depending on the confidence threshold value are depicted in Fig. 14 for TensorBox and Fig. 15 for Darknet-YOLO. As one can see, TensorBox localization performance does not depend on the confidence threshold value in any significant way. The best performance in terms of minimizing the number of false negative events with an acceptable number of false positive events can be achieved by maximizing the algorithm's accuracy metrics. For TensorBox, the maximum accuracy reaches a level of 31.6% for a confidence threshold value of zero with a corresponding polyp miss rate of 66.2%. For TensorBox, this is the best value, and it cannot be improved by adjusting the confidence threshold value. For Darknet-YOLO, maximum accuracy is reached at a 42.2% confidence

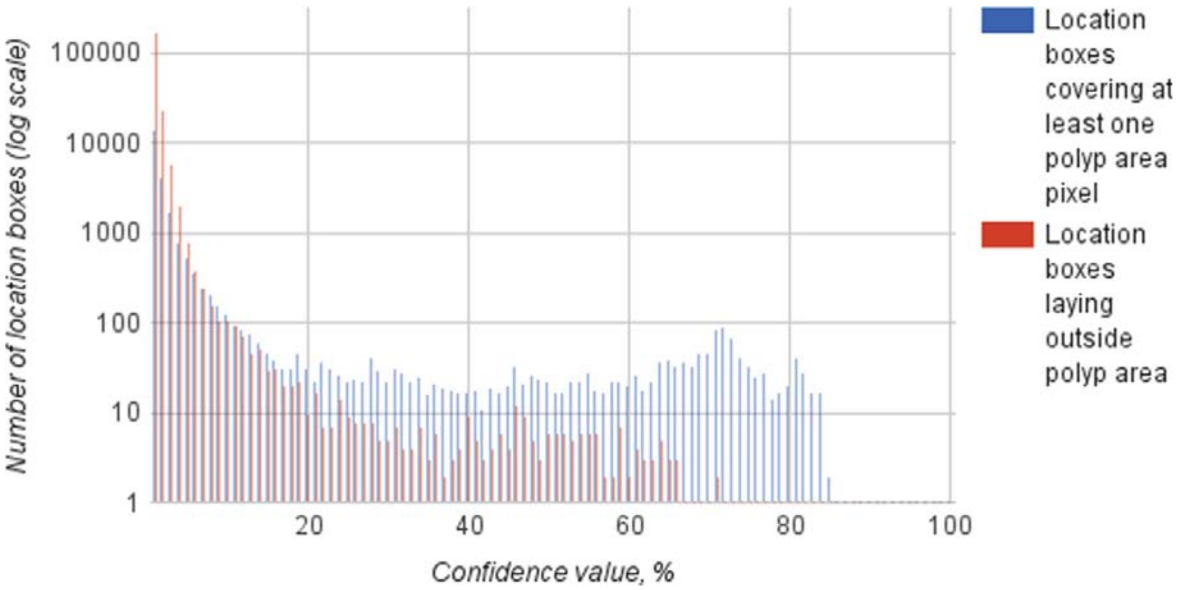


Fig. 13 The histogram shows confidence values and polyps area coverage by location boxes found by the Darknet-YOLO localization algorithm. It shows the expected high number of low-confident locations

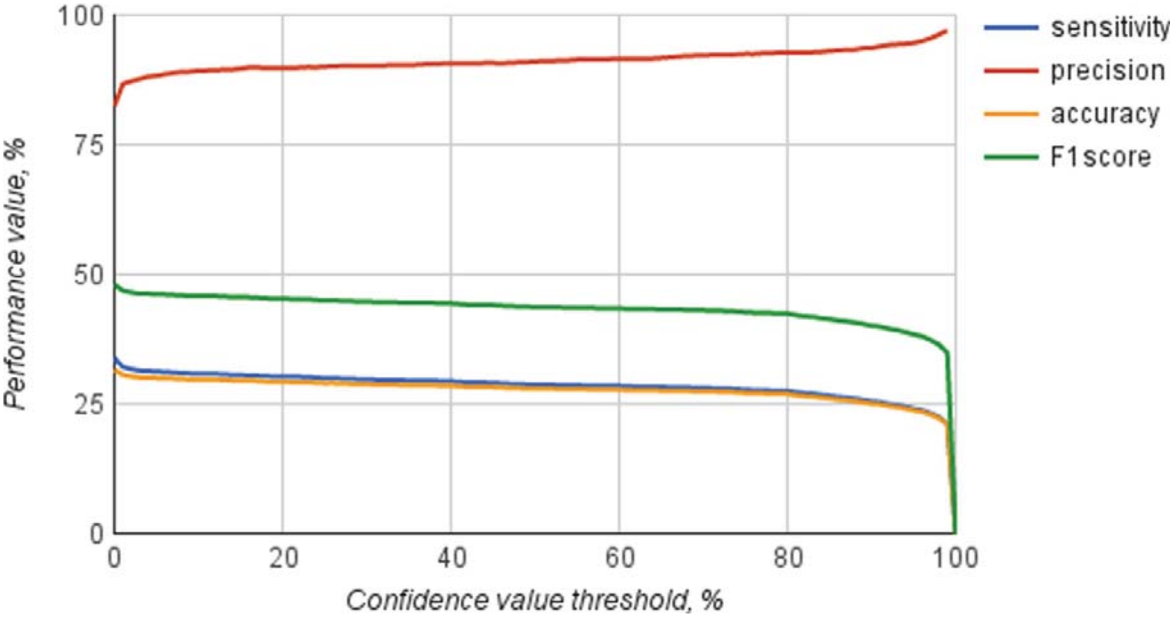


Fig. 14 The graphs show TensorBox localization algorithms performance for different confidence threshold values with no significant visible dependency. The maximum accuracy reaches level of 31.6% for zero-confidence threshold value with the polyp miss rate of 66.2%

threshold. The accuracy is 8% with a corresponding polyp miss rate of 47.9%. Darknet-YOLO showed more flexibility and a good response to the confidence threshold value. For Darknet-YOLO, the polyp miss rate can be significantly reduced by decreasing the confidence threshold value, but this gives a significant increase in the number of false positives, making the whole system too noisy. Nevertheless, combining Darknet-YOLO and the basic EIR localizer approaches can potentially give better overall system performance and better polyp miss rate.

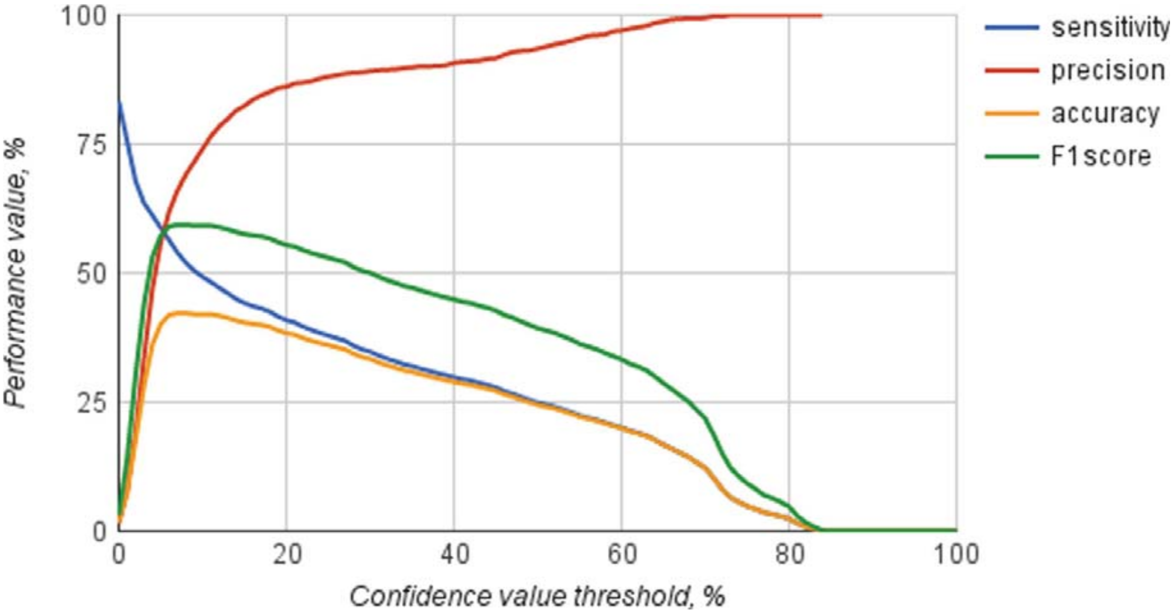


Fig. 15 The graphs show Darknet-YOLO localization algorithms performance for different confidence threshold values with good response to threshold value adjusting. The maximum accuracy reaches level of 42.2% for confidence threshold value of 8% with the polyp miss rate of 47.9%

Performing a comparison with well-known existing approaches in polyp localization is difficult due to lack of publicly available information (see Table 7) about other researchers' algorithms' performance and evaluation methods, and due to prevalent non-disclosure restrictions that prevent sharing of datasets in the research community. The available data shows, that our EIR basic localization approach has good performance with an F1 score of 41.6%.

The performance of the TensorBox approach (see Table 7) is too low for our real-time use-case. But, as depicted in Table 7, Darknet-YOLO performs well in terms of processing speed and can run at 45 frames per second. Our basic approach runs at 120 frames per second, thus a combination of both approaches can give us better localization performance while staying within the required real-time frame rate limits.

6 Real-world use cases

In this section, we describe two real-world use cases where the presented system can be used. The first one is a live system that will assist medical doctors during endoscopies. Currently, we are deploying a proof-of-applicabilty prototype in one of our partner hospitals. The second is a system that will automatically analyze videos captured by VCEs. Several hospitals are involved in this more concrete and applied research, and currently we are setting up the data-sharing agreements and collect the data for a new multi-disease dataset that will be released open-source. The first use case requires fast and reliable processing, and the second requires a system that is able to process a large amount of data in a reliable and scalable way.

6.1 Live system

The aim of the live system is to provide live feedback to the doctors, i.e., a computer-aided diagnosis in real-time. While the endoscopist performs the colonoscopy, the system analyzes the video frames that are captured by the colonoscope. To provide helpful information for the operating doctor, we combine the visual information from the endoscope with our marks. For the detection, we alter the frame borders and show the name of the detected finding in the auxiliary area of the endoscope device monitor. For the implemented localization classes, we put a cross on top of the localized findings (polyps in this system version). At the moment, we have implemented a demo version of the live system [39]. The live demo supports detection and localization of polyps. It is able to process a FullHD video stream with 30 FPS in real-time. An example of the graphical output of the live system is depicted in Fig. 16.

Table 7 Performance comparison of polyp localization approaches

System	True Pos.	False Pos.	False Neg.	Sensitivity	Precision	Accuracy	F1 score	FPS
Basic EIR	1266	3150	398	76.1%	28.7%	26.3%	**41.6%**	120
TensorBox-EIR	1459	311	2854	33.8%	82.4%	31.6%	**48.0%**	15
Darknet-YOLO-EIR	2245	1005	2068	52.1%	69.1%	42.2%	**59.4%**	43
Wang et al. [52]	–	–	–	95.7%	–	–	**95.7%**	10
Hwang et al. [18]	–	–	–	96.0%	83.0%	–	**–**	15

Bold numbers shows the balanced F-score of each proposed method

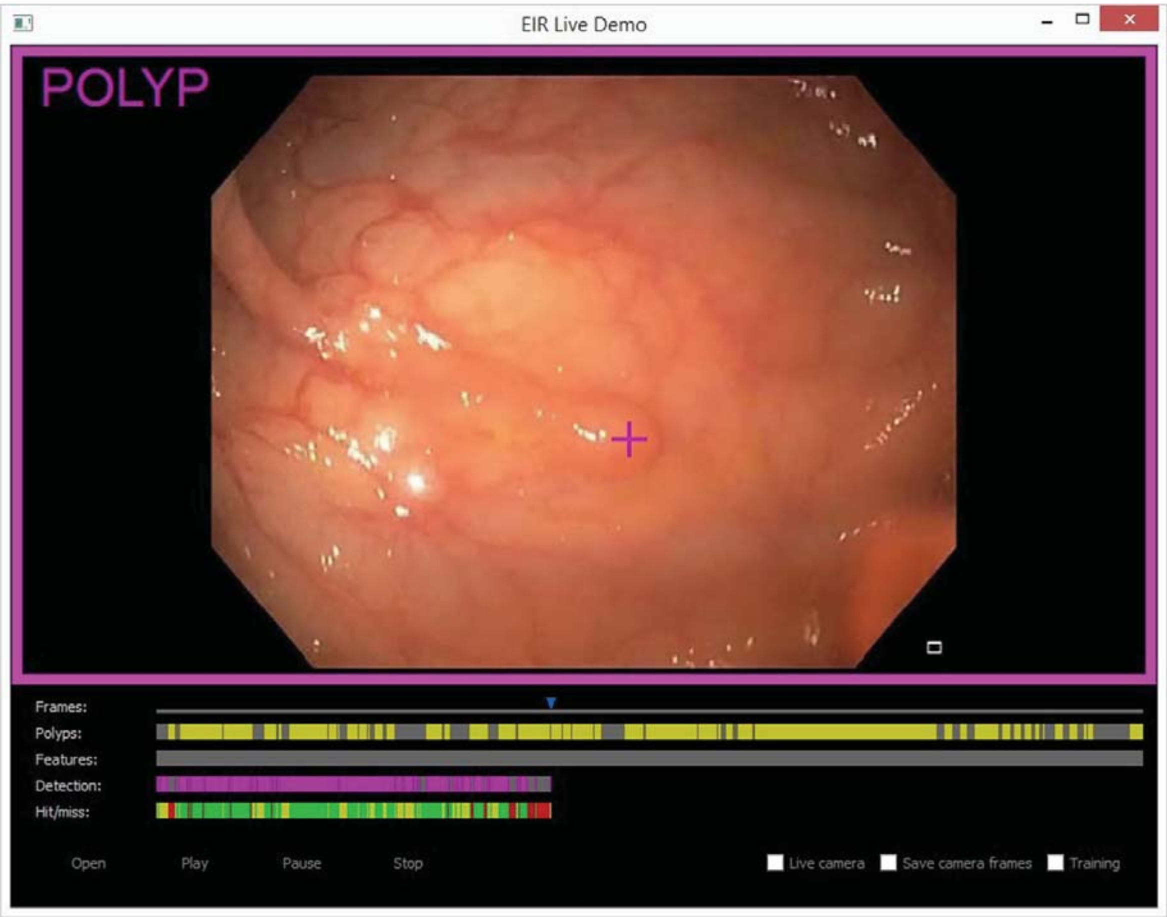


Fig. 16 A screenshot of the live system showing the combination of the visual information from the endoscope with feedback information from the detection and localization system. The pink fame surrounding background shows a positive detection. The name of the detected finding is shown in the frame auxiliary screen area, and the cross shows the location of the polyp

In addition to supporting the medical expert during the colonoscopy, we are working on an extension of the system, where the system is used to document the examination procedure. We will implement the generation of a document with an overview of the colonoscopic procedure. The doctors will be able to make changes or corrections, and add additional information to that document. The document will be stored or used as an appendix to the written endoscopy report.

6.2 Wireless video capsule endoscope

The current existing VCEs have a resolution of around 256 × 256, frame rates of 3-35 frames per second (adaptive frame rate with a feedback loop from the receiver to the transmitter). They do not have optimum lighting, making it more difficult to detect endoscopic findings in the captured images than in images from traditional endoscopes. Also, during VCE procedures, the intestine is not expanded, unlike in traditional endoscopy, where the expansion allows for clear and non-obfuscated pictures of the intestine walls. Nevertheless, ongoing research aims at improving the VCEs' hardware capabilities and at improving the methods and algorithms developed for colonoscopies to work also for VCEs [22]. The multi-sensor VCE is swallowed in order to visualize the GI tract for subsequent diagnosis and detection of GI diseases. Thus, people may in the future be able to buy VCEs at the pharmacy, and deliver the video stream from the GI tract to the phone over a wireless connection. In the best case, the first screening results are available within eight hours after swallowing the VCE, which is the time the camera typically spends traversing the GI tract. Thus, the ability to implement and perform mass-screening of the GI tract highly depends on two main research areas. First, it requires the development of a new generation of VCEs with better image quality and the ability to communicate with widely used mobile phones. Second, mass-screening requires a new generation of lesion detection algorithms able to process the captured GI tract multimedia data and video footage fully automatically in the mobile phone with public cloud computing support. Here, a preliminary analysis and task-oriented compression of a captured video footage before uploading into the cloud is important due to huge amounts of video data generates by VCEs. In our future research for this use case, we will work on the adaptation of the detection algorithms to the common mobile platforms. We will create a new mobile application to demonstrate the ability of our system to perform on hardware with the limited resources available.

7 Conclusion

In this paper, a complex automated diagnosis system built for different GI tract disease detection scenarios, colonic polyp localization and big dataset visualization has been presented. We briefly described the whole system from data collection for medical knowledge transfer and system learning, evaluation of the experimental results to visualization of the findings. A detailed evaluation of detection of multiple endoscopic findings, polyp-localization accuracy and system performance has been performed. We introduced two new multi-class classification methods, one based on a deep learning neural network approach and another new multi-class classification algorithm based on global image features. For the localization, we evaluated existing localization approaches based on deep learning neural networks and compared the results to our initial localization method.

The novelty of the research includes an end-to-end implementation of the whole EIR system pipeline, from frame capture, annotation and analysis to user (doctor) feedback, as a combination of many out-of-the-box and modified existing components, as well as several new ones. The experiments showed that the proposed system (i.e., both the global feature-based and the neural network-based implementations) can achieve equal results to state-of-the-art methods in terms of detection performance for state-of-the-art endoscopic data, and a comparable localization performance. Further, we showed that the new EIR system outperforms state-of-the-art systems in terms of system performance, that it scales in terms of data throughput and that it can be used in a real-time scenario. We concluded, based on our initial experiments, that the global features multi-class detection approach slightly outperforms the tested neural network approaches, and that the localization algorithm can benefit from combining local features and neural network approaches. We also presented automatic analysis of VCE videos and live support of colonoscopies as two real-world use cases that can potentially benefit from the proposed system where clinical tests are currently being planned in our partner hospitals. The experimental evaluation of the system as well as dataset creation are performed in collaboration with the Cancer Registry of Norway, and in the near future, the system will be tested in a real-world environment, i.e., it will have a real societal impact.

For future work, we plan to further improve the multi-class detection and localization accuracy of the system and support detection and localization of more abnormalities. In this respect, we are currently working with medical experts to collect more training data, annotate them and create new, larger training and testing datasets [30, 31]. Finally, to further improve the performance of the system, we work on a universal system extension that will allow the system to utilize the computing power of one or more GPUs on single or multiple nodes. Implementing such an extension will allow parallelization of the detection and localization workloads [32], which is important in our multi-disease analysis system of GI tract [32, 35, 37–39].

Acknowledgements This work is founded by the FRINATEK project “EONS” #231687.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

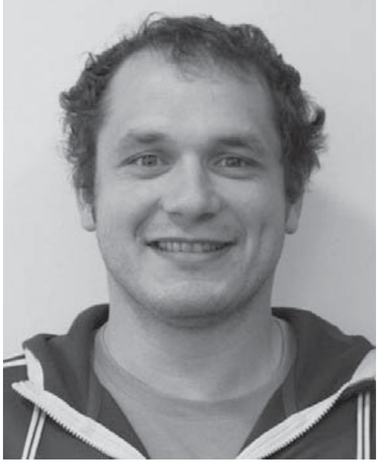
References

1. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, etal (2016) Tensorflow: a system for large-scale machine learning. In: Proceedings of OSDI
2. Albisser Z, Riegler M, Halvorsen P, Zhou J, Griwodz C, Balasingham I, Gurrin C (2015) Expert driven semi-supervised elucidation tool for medical endoscopic videos. In: Proceedings of MMSys, pp 73–76
3. Alexandre LA, Casteleiro J, Nobreinst N (2007) Polyp detection in endoscopic video using svms. In: Proceedings of PKDD, pp 358–365
4. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46(3):175–185
5. Ameling S, Wirth S, Paulus D, Lacey G, Vilarino F (2009) Texture-based polyp detection in colonoscopy. In: Proceedings of bfm, pp 346–350
6. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
7. Buades A, Coll B, Morel JM (2011) Non-local means denoising. Image Processing On Line 1:208–212
8. Chaabouni S, Benois-Pineau J, Amar CB (2016) Transfer learning with deep networks for saliency prediction in natural video. In: Proceedings of ICIP, pp 1604–1608
9. Chatzichristofis S, Boutalis Y (2008) Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. Comput Vis Syst 312–322
10. Chatzichristofis SA, Boutalis YS (2008) Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In: 9th international workshop on image analysis for multimedia interactive services, 2008. WIAMIS’08. IEEE, pp 191–196
11. Cheng DC, Ting WC, Chen YF, Pu Q, Jiang X (2008) Colorectal polyps detection using texture features and support vector machine. In: Proceedigns of MDAISM, pp 62–72
12. Chin C, Brown DE (2000) Learning in science: a comparison of deep and surface approaches. J Res Sci Teach 37(2):109–138
13. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Proceedings of CVPR, pp 248–255
14. Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) Decaf: a deep convolutional activation feature for generic visual recognition. In: Proceedings of ICML, pp 647–655
15. Fitzgibbon AW, Fisher RB et al (1996) A buyer’s xguide to conic fitting. DAI Research paper
16. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. ACM SIGKDD Explor Newslet 11(1):10–18
17. Holme Ø., Bretthauer M, Fretheim A, Odgaard-Jensen J, Hoff G (2013) Flexible sigmoidoscopy versus faecal occult blood testing for colorectal cancer screening in asymptomatic individuals. The Cochrane Library
18. Hwang S, Oh J, Tavanapong W, Wong J, de Groen P (2007) Polyp detection in colonoscopy video using elliptical shape feature. In: Proceedings of ICIP, pp 465–468
19. Imagenet ImageNet Challenge Datasets. http://www.image-net.org/. [last visited, March 06, 2016]
20. Kaminski MF, Regula J, Kraszewska E, Polkowski M, Wojciechowska U, Didkowska J, Zwierko M, Rupinski M, Nowacki MP, Butruk E (2010) Quality indicators for colonoscopy and the risk of interval cancer. N Engl J Med 362(19):1795–1803

21. Kang J, Doraiswami R (2003) Real-time image processing system for endoscopic applications. In: Proceedings of CCECE, vol 3, pp 1469–1472
22. Khaleghi A, Balasingham I (2015) Wireless communication link for capsule endoscope at 600 mhz. In: Proceedings of EMBC, pp 4081–4084
23. Li B, Meng MH (2012) Tumor recognition in wireless capsule endoscopy images using textural features and svm-based feature selection. IEEE Trans Inf Technol Biomed 16(3):323–329
24. Lux M, Marques O (2013) Visual information retrieval using java and lire. Synt Lect Inform Conc Retri Serv 5(1):1–112
25. Mallery S, Van Dam J (2000) Advances in diagnostic and therapeutic endoscopy. Med Clin N Am 84(5):1059–1083
26. Mamonov A, Figueiredo I, Figueiredo P, Tsai YH (2014) Automated polyp detection in colon capsule endoscopy. IEEE Trans Med Imaging 33(7):1488–1502
27. Ngiam J, Coates A, Lahiri A, Prochnow B, Le QV, Ng AY (2011) On optimization methods for deep learning. In: Proceedings of ICML, pp 265–272
28. Nguyen A, Yosinski J, Clune J (2014) Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. arXiv:1412.1897
29. O'Connell JB, Maggard MA, Ko CY (2004) Colon cancer survival rates with the new american joint committee on cancer sixth edition staging. J Natl Cancer Inst 96(19):1420–1425
30. Pogorelov K, Randel KR, Griwodz C, Eskeland SL, de Lange T, Johansen D, Spampinato C, Dang-Nguyen DT, Lux M, Schmidt PT, Riegler M, Halvorsen P (2017) Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of MMSYS, pp 164–169
31. Pogorelov K, Randel KR, de Lange T, Eskeland SL, Griwodz C, Johansen D, Spampinato C, Taschwer M, Lux M, Schmidt PT, Riegler M, Halvorsen P (2017) Nerthus: a bowel preparation quality video dataset. In: Proceedings of MMSYS, pp 170–174
32. Pogorelov K, Riegler M, Halvorsen P, Schmidt PT, Griwodz C, Johansen D, Eskeland SL, de Lange T (2016) GPU-Accelerated real-time gastrointestinal diseases detection. In: Proceedings of CBMS, pp 185–190
33. Redmon J Darknet: Open source neural networks in C. http://pjreddie.com/darknet/. [last visited, March 06, 2016]
34. Redmon J, Divvala S, Girshick R, Farhadi A (2015) You only look once: Unified, real-time object detection. arXiv:1506.02640
35. Riegler M, Griwodz C, Spampinato C, de Lange T, Eskeland SL, Pogorelov K, Tavanapong W, Schmidt PT, Gurrin C, Johansen D, Johansen H, Halvorsen P (2016) Multimedia and medicine: Teammates for better disease detection and survival. In: Proceedings of ACM MM, pp 968–977
36. Riegler M, Pogorelov K, Eskeland SL, Thelin Schmidt P, Albisser Z, Johansen D, Griwodz C, Halvorsen P, de Lange T (2017) From annotation to computer aided diagnosis: Detailed evaluation of a medical multimedia system. ACM Trans Multimed Comput Commun Appl 9(4)
37. Riegler M, Pogorelov K, Halvorsen P, de Lange T, Griwodz C, Johansen D, Schmidt PT, Eskeland SL (2016) Eir - efficient computer aided diagnosis framework for gastrointestinal endoscopies. In: Proceedings of CBMI, pp 1–6
38. Riegler M, Pogorelov K, Lux M, Halvorsen P, Griwodz C, de Lange T, Eskeland SL (2016) Explorative hyperbolic-tree-based clustering tool for unsupervised knowledge discovery. In: Proceedings of CBMI, pp 1–4
39. Riegler M, Pogorelov K, Markussen J, Lux M, Stensland HK, de Lange T, Griwodz C, Halvorsen P, Johansen D, Schmidt PT, Eskeland SL (2016) Computer aided disease detection system for gastrointestinal examinations. In: Proceedings of MMSys, p 29
40. Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117
41. Stewart R, Andriluka M (2015) End-to-end people detection in crowded scenes. arXiv
42. Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2014) Going deeper with convolutions. CoRR 1409.4842
43. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the inception architecture for computer vision. arXiv:1512.00567
44. Tajbakhsh N, Gurudu SR, Liang J (2016) Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imaging 35(2):630–644
45. Tamura H, Mori S, Yamawaki T (1978) Textural features corresponding to visual perception. IEEE Trans Syst Man Cybern 8(6):460–473

46. Tanimoto TT (1958) Elementary mathematical theory of classification and prediction
47. The New York Times: The 2.7 Trillion Medical Bill. http://goo.gl/CuFyFJ. [last visited, Nov. 29, 2015]
48. Tieleman T, Hinton G (2012) Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude COURSERA. Neural Networks for Machine Learning 4(2)
49. Van Essen B, Macaraeg C, Gokhale M, Prenger R (2012) Accelerating a random forest classifier: Multi-core, GP-GPU, or FPGA? In: Proceedings of FCCM, pp 232–239
50. von Karsa L, Patnick J, Segnan N (2012) European guidelines for quality assurance in colorectal cancer screening and diagnosis. first edition–executive summary. Endoscopy 44(S 03):SE1–SE8
51. Wang Y, Tavanapong W, Wong J, Oh J, de Groen PC (2014) Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy. In: Proceedings of BHI, vol 18, pp 1379–1389
52. Wang Y, Tavanapong W, Wong J, Oh JH, De Groen PC (2015) Polyp-alert: Near real-time feedback during colonoscopy. Comput Meth Programs Biomed 120(3):164–179
53. Zagoris K, Chatzichristofis SA, Papamarkos N, Boutalis YS (2010) Automatic image annotation and retrieval using the joint composite descriptor. In: 14th panhellenic conference on informatics (PCI), 2010. IEEE, pp 143–147
54. Zhou M, Bao G, Geng Y, Alkandari B, Li X (2014) Polyp detection and radius measurement in small intestine using video capsule endoscopy. In: Proceedings of BMEI, pp 237–241

Konstantin Pogorelov



Michael Riegler

Sigrun Losada Eskeland

Thomas de Lange

Dag Johansen

Carsten Griwodz

Peter Thelin Schmidt

Pål Halvorsen