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

Konstantin Pogorelov

17.07.2019

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

List of Figures

List of Tables

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 [116] 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 [116] 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 [147]. 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 [143]. Several screening methods exist, e.g., fecal immunochemical tests (FITs), sigmoidoscopy screening, computer tomography (CT)

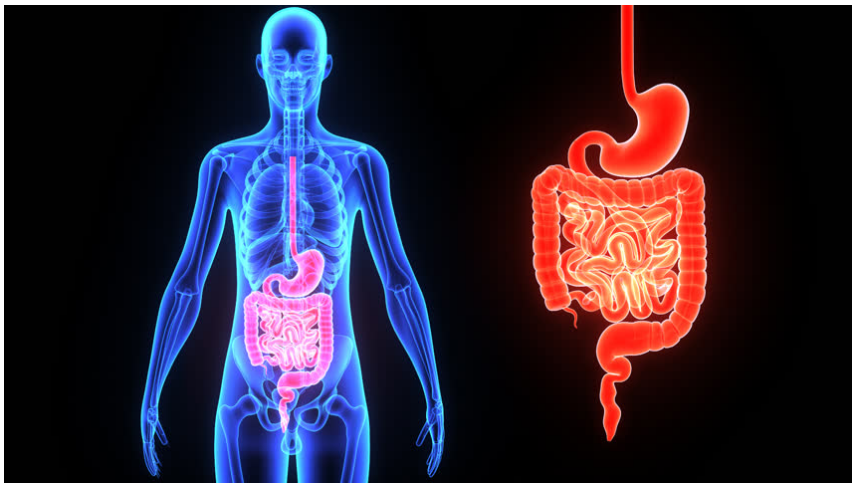
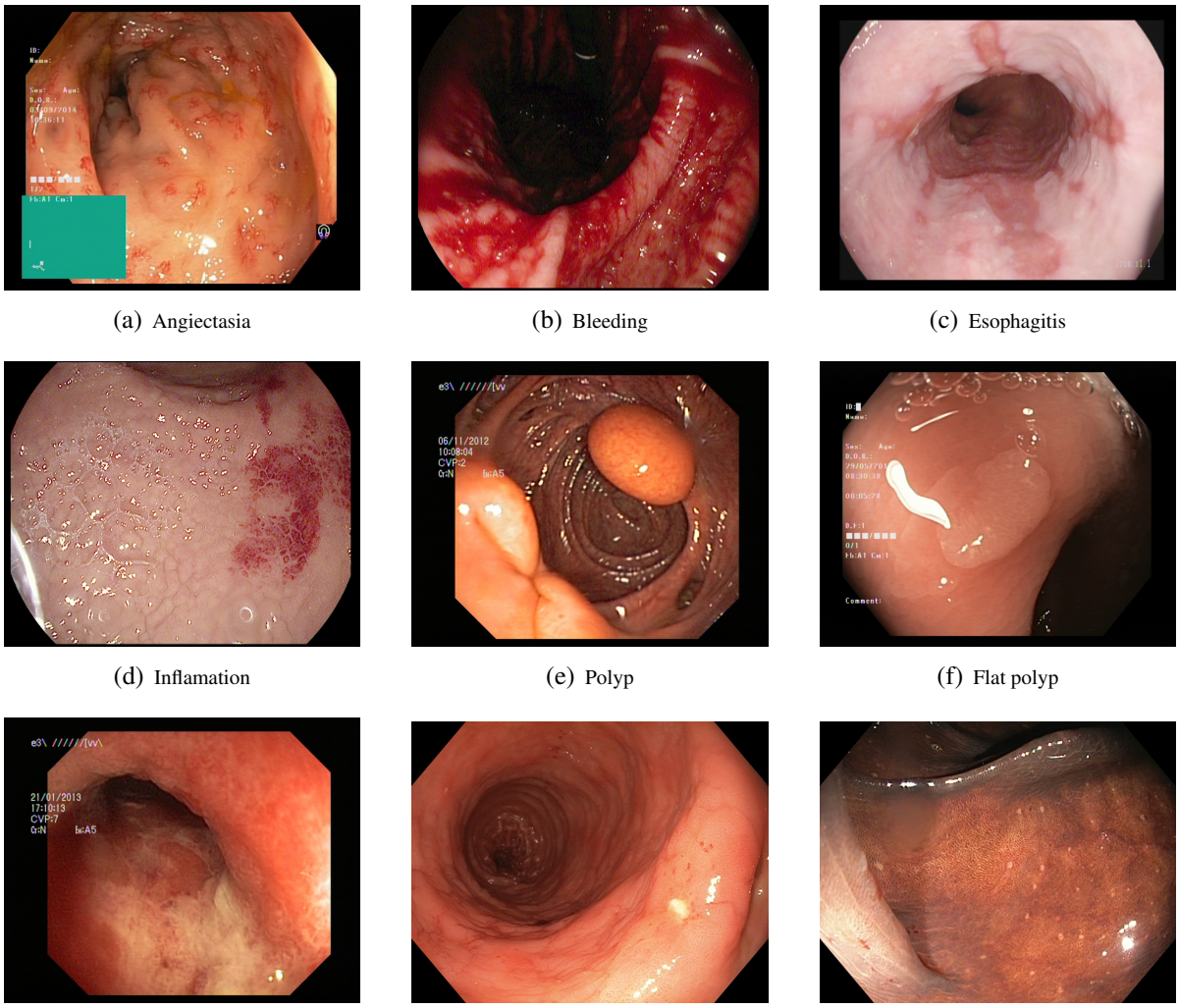


Figure 1.1: An overview of the human GI tract (hdfootagestock.com).



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

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

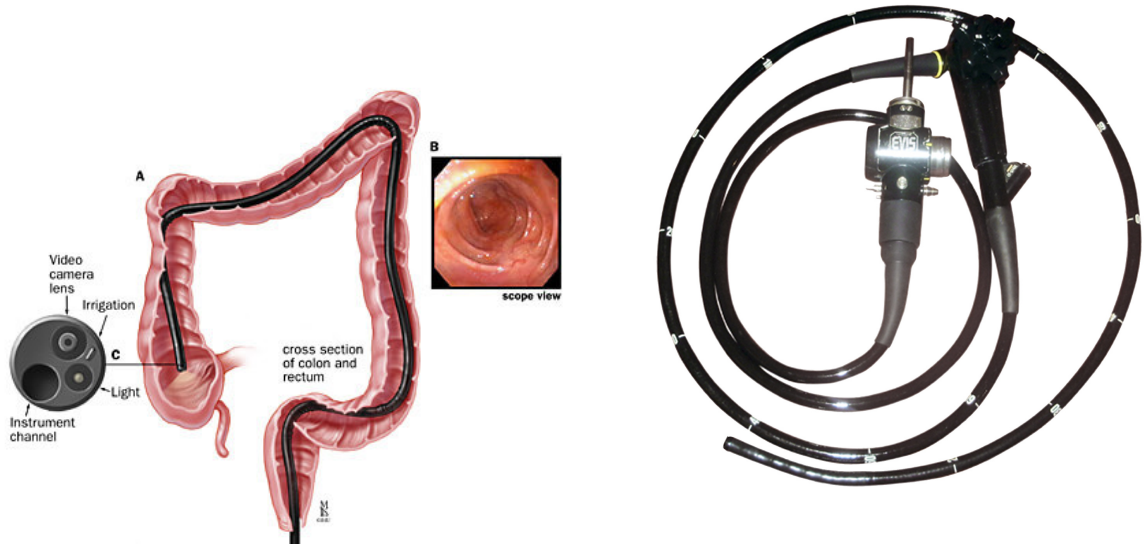
(g) Ulcerative colitis (h) Erosion (i) Melanosis

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 [142] for



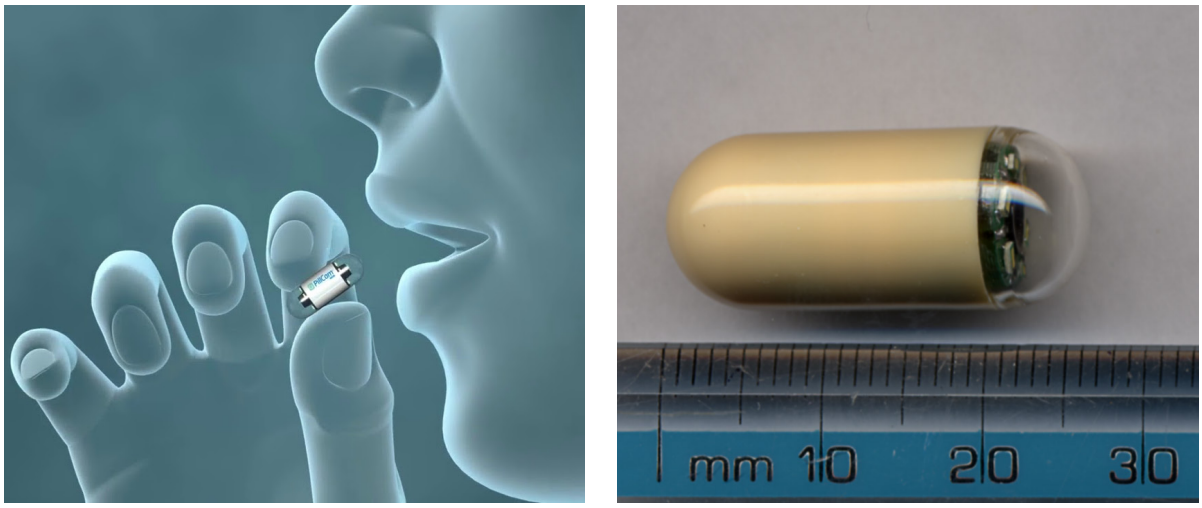
(a) Colonoscopy (hopkinsmedicine.org) (b) Coloscope (olympus.com)

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) [115].

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 [136], i.e., an average of $1,100 per examination [137] (up to $6,000 in New York). In the United Kingdom, the costs are around $2,700 per examination [122]. 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 [115] 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 [115].

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 [115] 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:

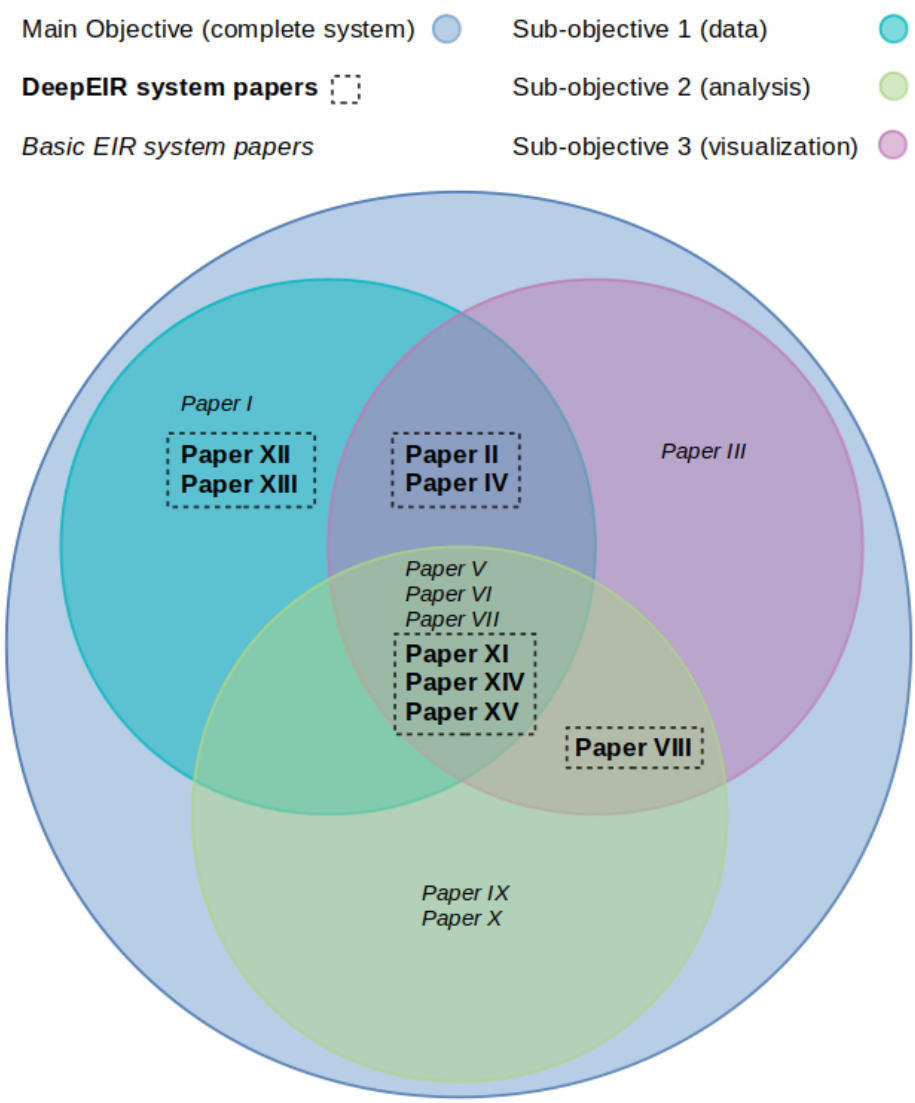


Figure 1.5: This diagram depicts the contributions for each of the in part II attached papers to the, for this thesis defined, objectives.

- **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, 116, 117, 120].

  Using the ASU Mayo dataset [133], 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, 116, 117, 120].

  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, 116, 120]. 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 [116]. 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, 120]. 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 [125]. 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, 116, 120]. All-in-all, we contributed with a holistic medical multimedia system for GI examinations [115] 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, 119].

  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, 118]. 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 [128].

  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, 115, 116, 117, 120].

- **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 [120]. We developed an initial visualization approach that is able to visualize all outputs of the DeepEIR system [116], 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, 121] 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 [114].
- 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 [128]. 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 pro-

vide 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.

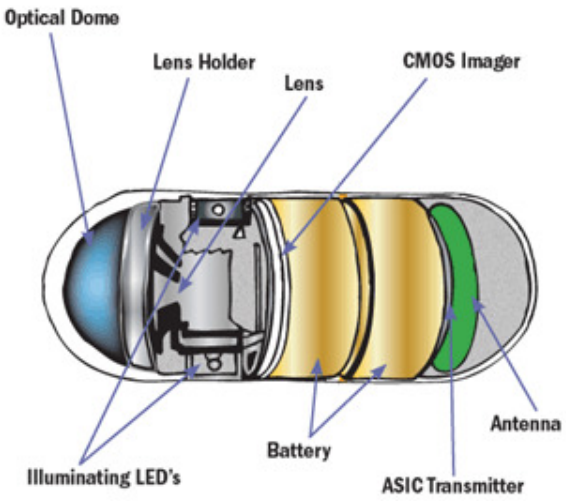


Figure 2.1: The internal components of wireless video capsule endoscope

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. [144, 145] 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.[139]	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. [145]	2015	polyp / edge, texture	97.70	–	–	95.70	–	–	10	1.8m
Wang et al. [144]	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. [149]	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 [115]. 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. [139] 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.

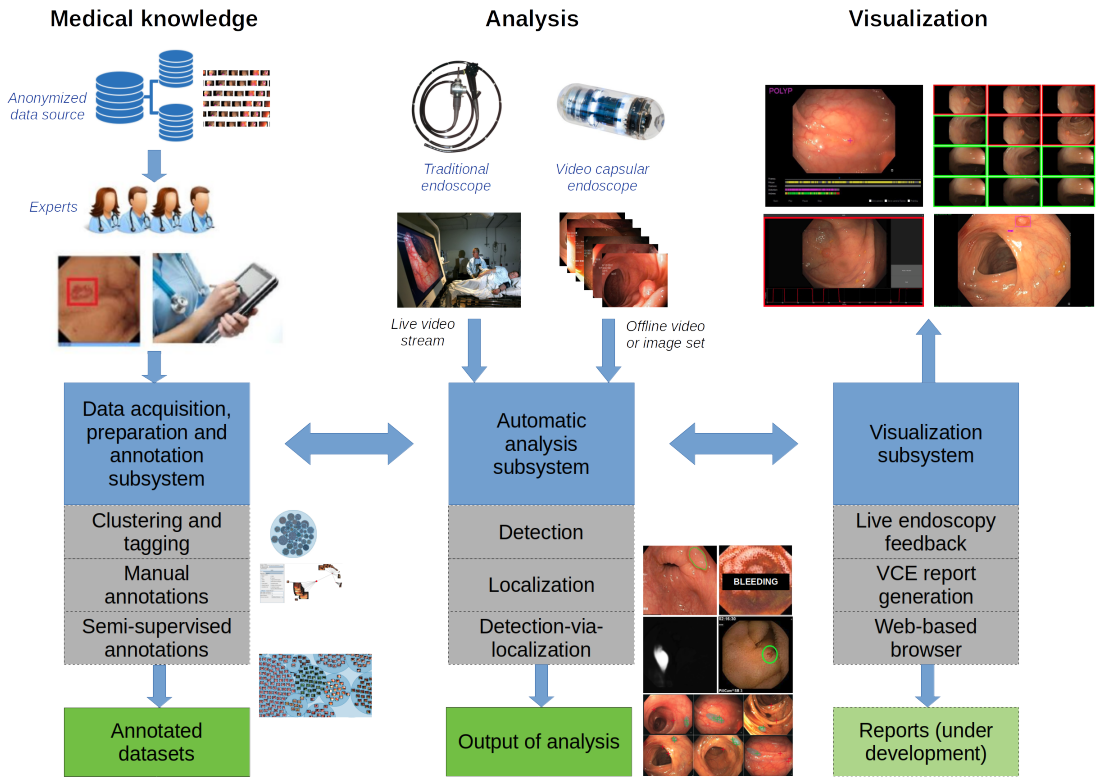


Figure 3.1: A complete overview of the DeepEIR system. The system consists of data acquisition, preparation and annotation, automatic analysis and visualization subsystems.

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 sub-versions. 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.

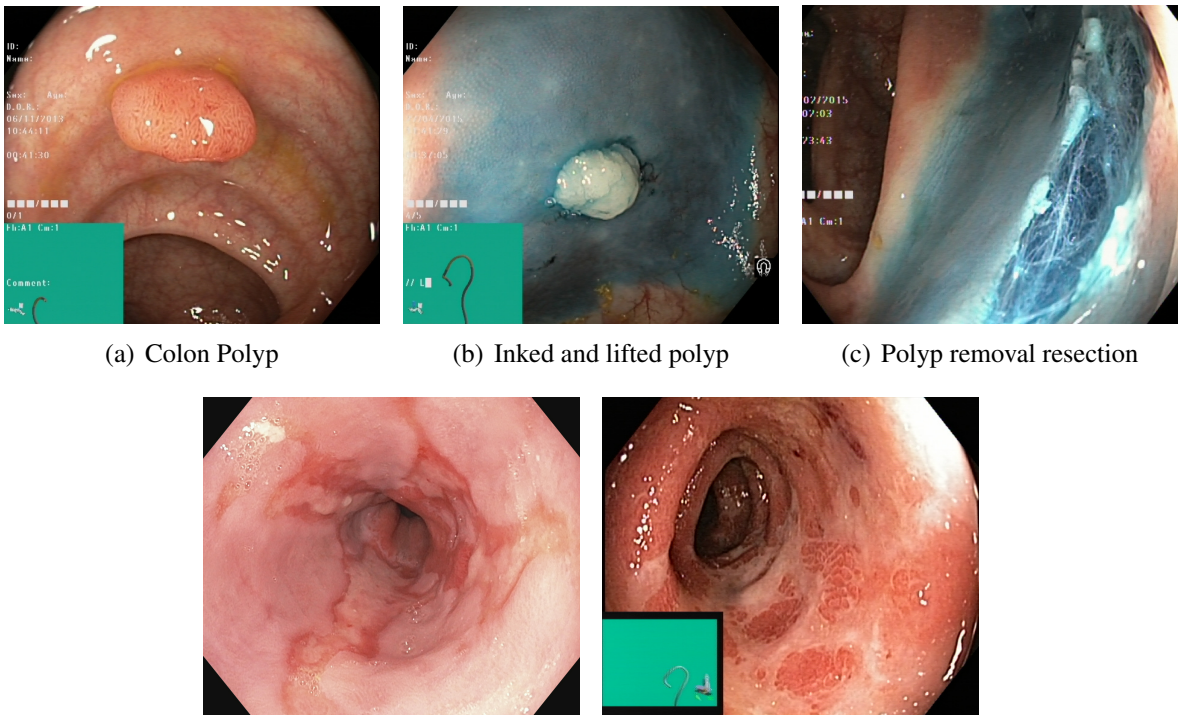
The after-removal resection margins are important in order to evaluate whether the polyp is completely removed or not. Residual polyp tissue may lead to continued growth and in the worst case malignancy development. Figure 3.2(c) illustrates the resection site after removal of a polyp. Automatic recognition 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

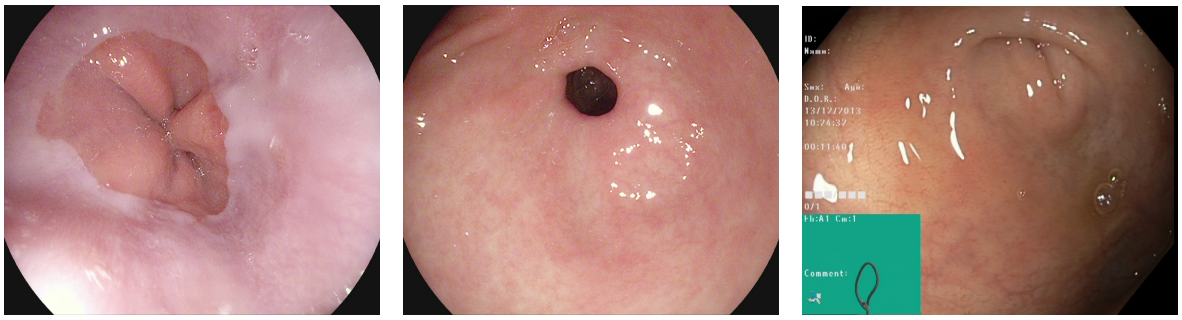
An anatomical landmark is a recognizable feature within the GI tract that is easily visible through the endoscope. Landmarks are essential for navigation and as a reference point for describing the location of a given finding. The landmarks are also be typical sites for pathology like ulcers or inflammation. 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.

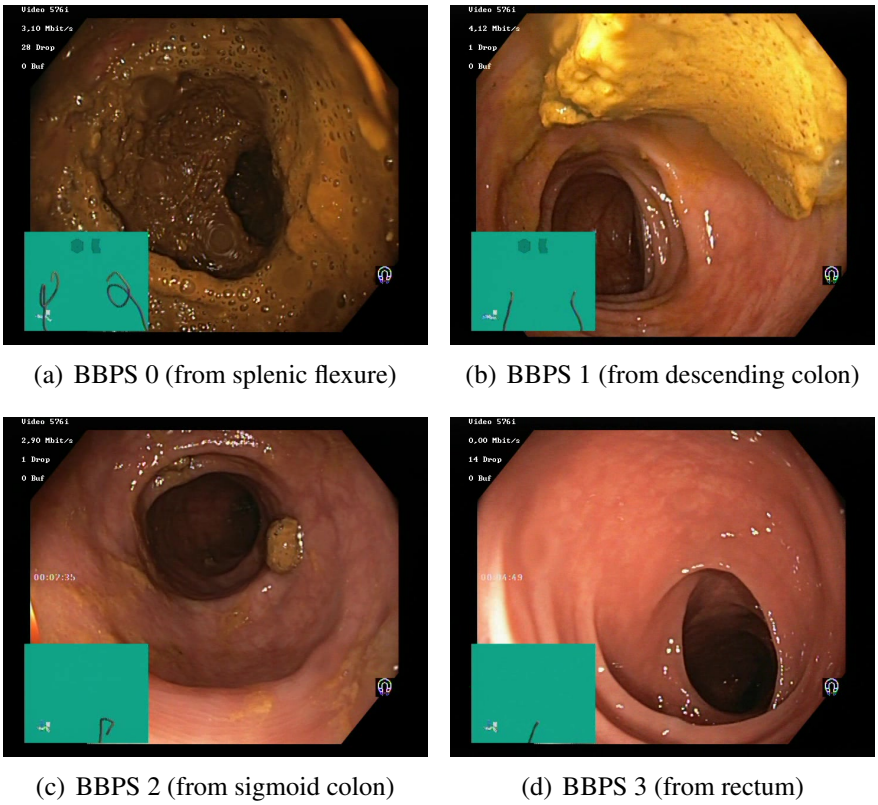
Cecum

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, 140]. 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) [148], Tamura [134], 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".

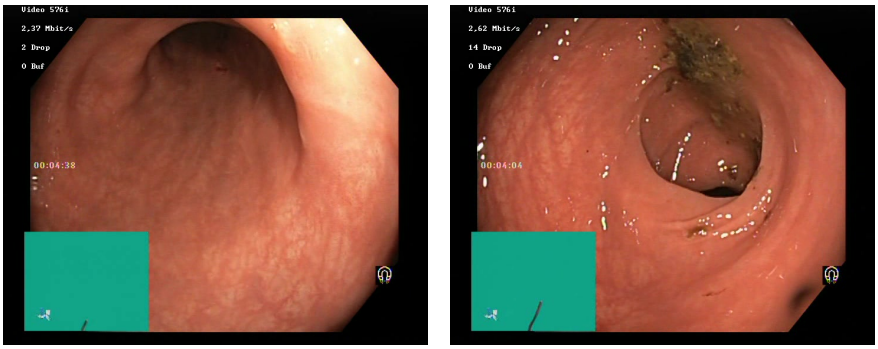
In total, the Medico dataset includes 16 classes showing anatomical landmarks, phatological findings or endoscopic procedures in the GI tract. The anatomical landmarks are Z-line, pylorus and cecum, while the pathological findings include esophagitis, polyps and ulcerative colitis. In addition, we provide 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

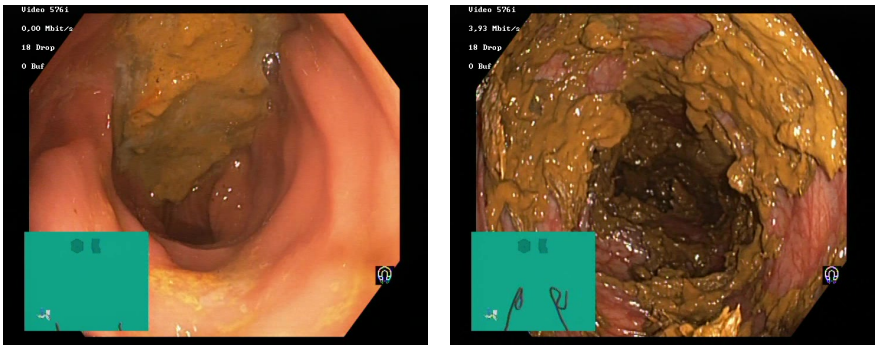
This class represents the samples of normal tissue that can be observed during colonoscopies (see figure 3.5(a) for an example). Comparing to abnormalities, there is no interest in detecting this type of image during live colonoscopies. However, we think that this class can be used for the opposite detection 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.

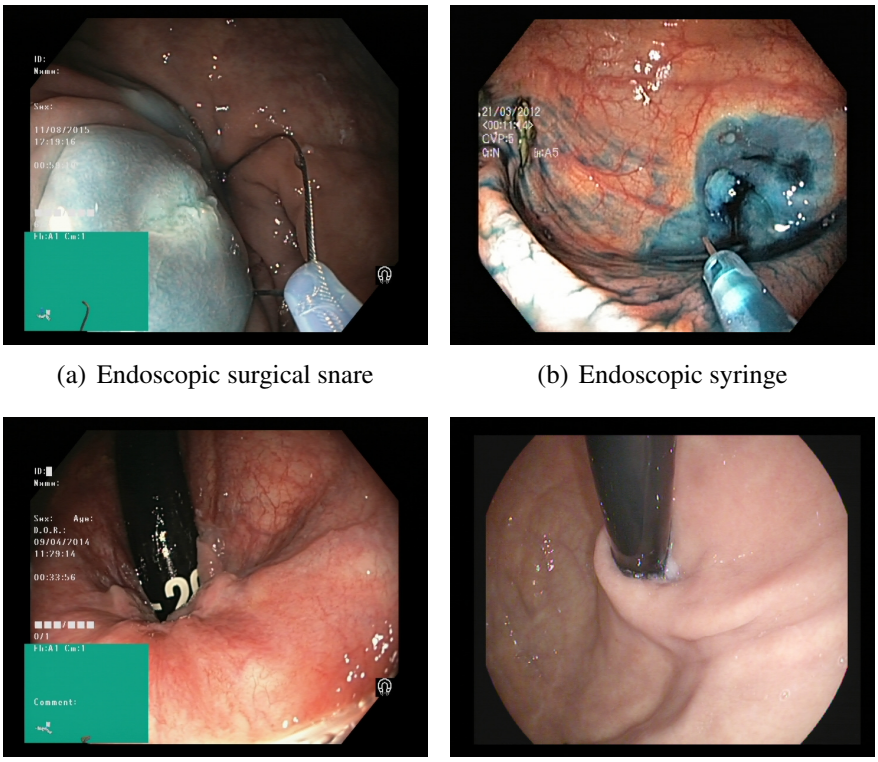
accurate interpretation of the bowel cleanliness is important and, therefore, we added a two new classes, both containing different amounts of stool masses (see figure 3.5 for the examples of stool inclusions 3.5(b), medium 3.5(c) and significant 3.5(d) amounts of stool).

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).

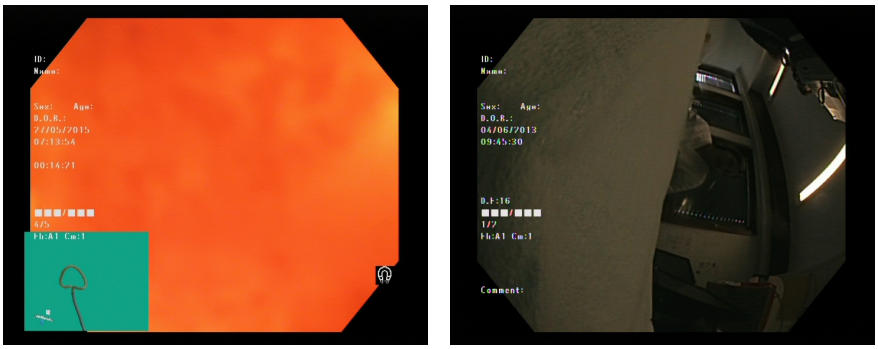
endoscopic procedure (see figure 3.7). Out-of-patient images can also be used for detection the begin and end of endoscopic procedure, which is important for automated reporting generation.

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.

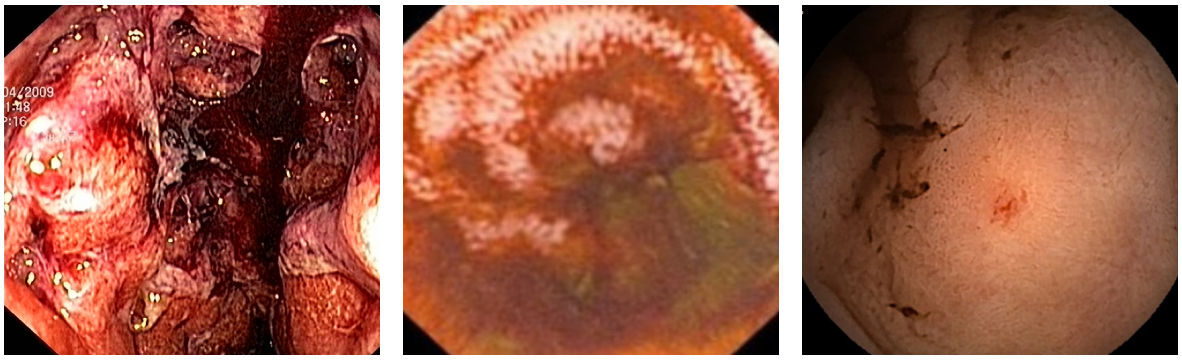
Colorectal Cancer

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 [115]. 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) [148], Fuzzy Color and Texture Histogram (FCTH) [38], Tamura [134], 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/`

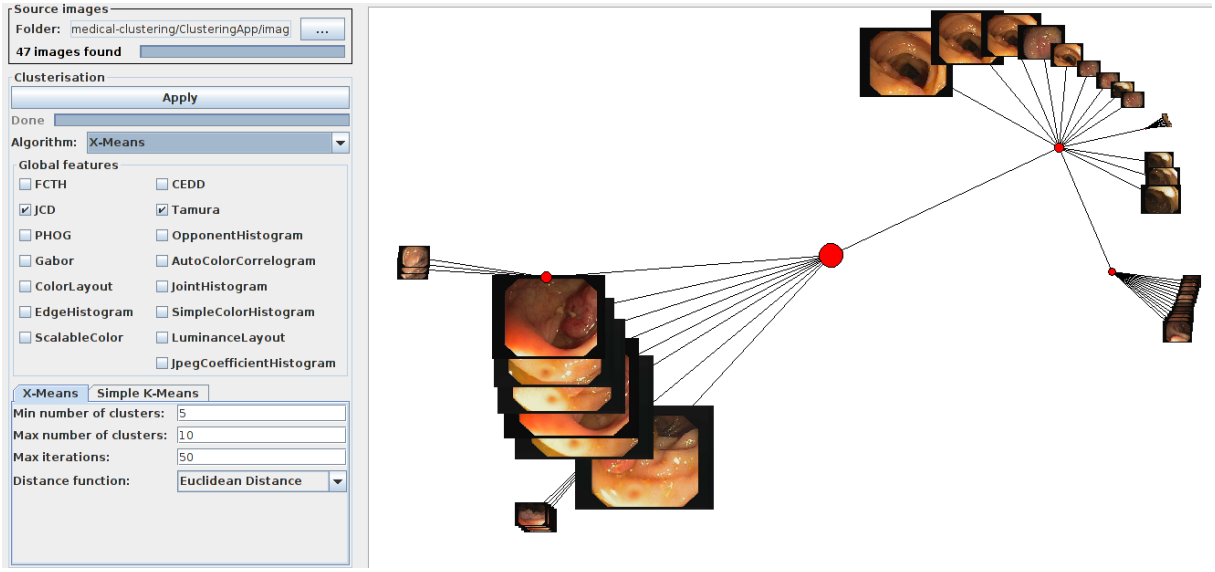


Figure 3.9: Hyper-tree based visualization, clustering and annotation system.

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.

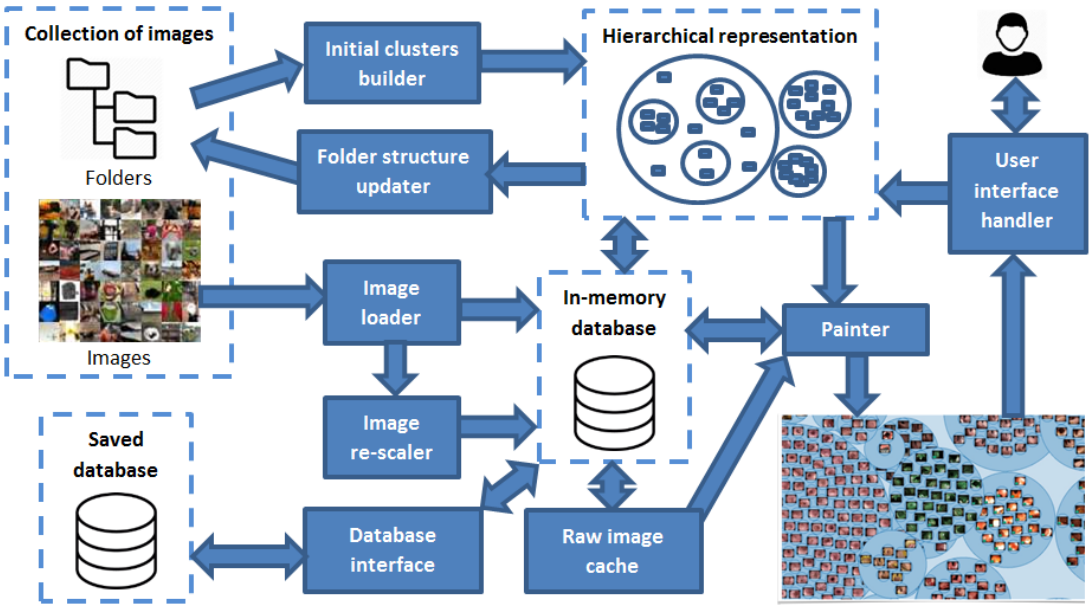


Figure 3.10: Structure of the visualization and user interface engine of the presented ClusterTag application. A number of caching and intermediate data processing routines are used to make it possible to perform real-time visualization and interaction with huge image collections.

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

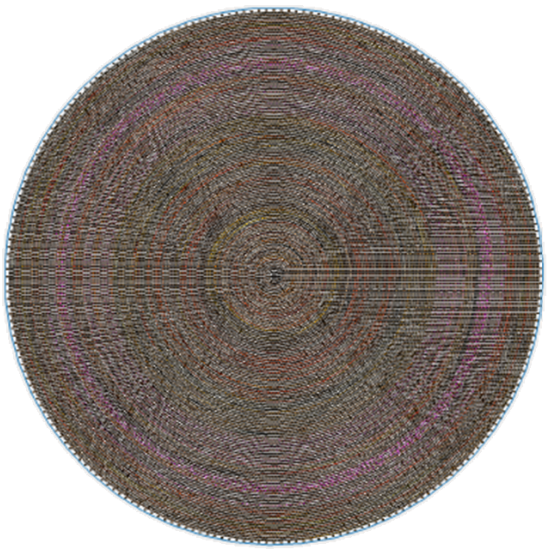
optimized visualization engine written in Java. The overall structure of the software is depicted in figure 3.10. It consist of the following sub-modules:

- 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.

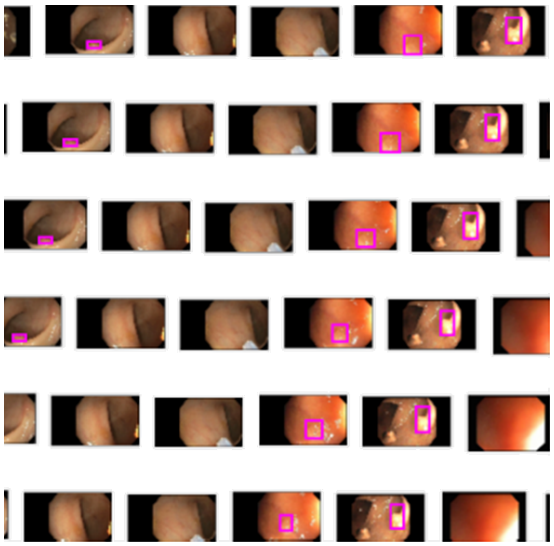
We have designed and implemented several additional optimization techniques to allow real-time handling of huge image collections. The most important are a database of ready-to-draw pre-processed images, caching of raw image visual representation, painting of adaptive image spatial resolution, interaction 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.

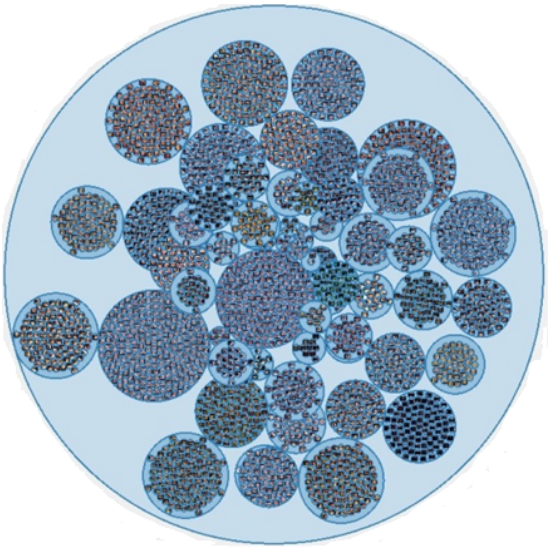
The ClusterTag tool was intensively used during the Kvasir and Nerthus dataset creation and annotation. It already demonstrated a great potential for big image processing and was evaluated with different end-users and domain experts including experienced medical doctors.



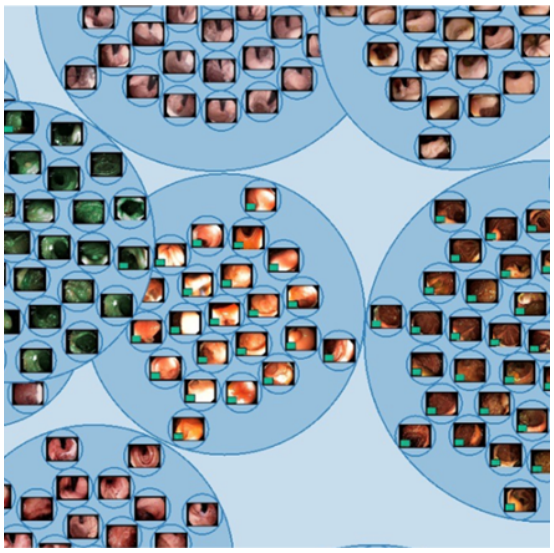
(a) Unsorted collection of images.



(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, 116, 117] 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 [114] 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`

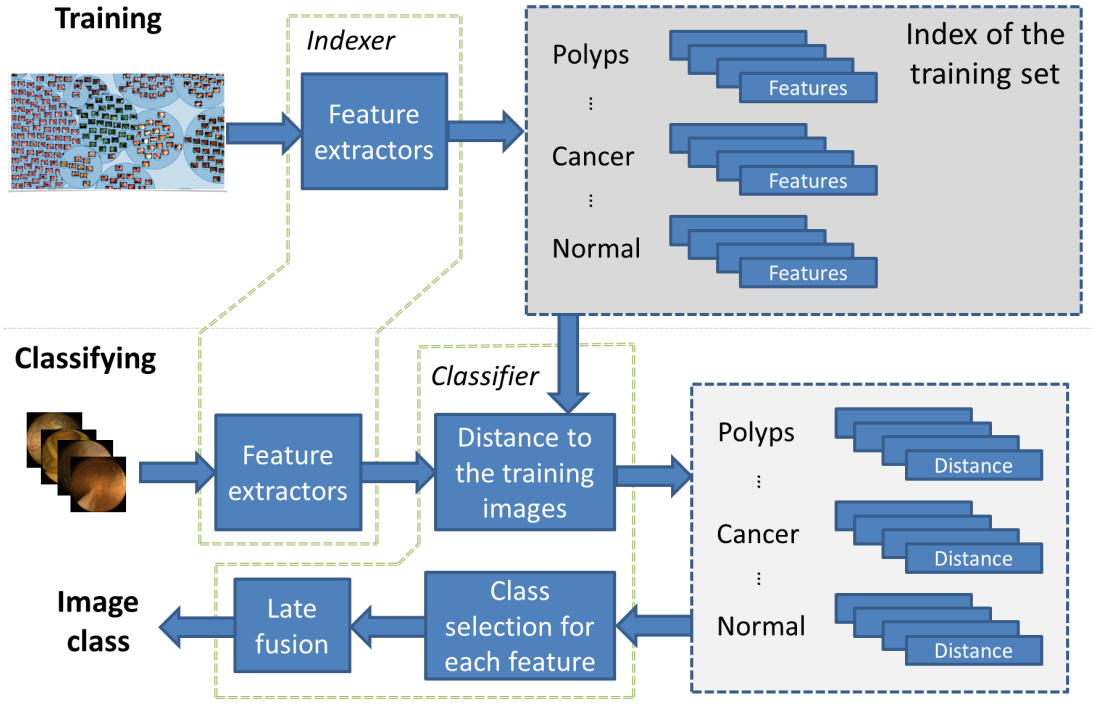


Figure 3.12: Detailed steps for the multi-class global-feature-based detection implementation

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, 117]. 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 [135]. A smaller distance between an image in the index and the query image means a better rank [135]. The final ranked list is used in the classification step, which imple-

ments 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 [141] 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 [131] 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 [131] 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 [138] 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.

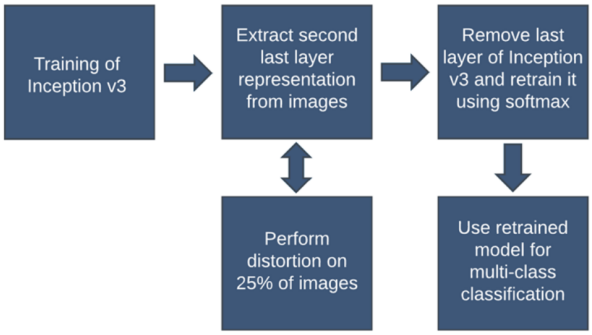


Figure 3.13: Multi-class deep-learning-based detection pipeline

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 [124], InceptionV3 [132] 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 [129] 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

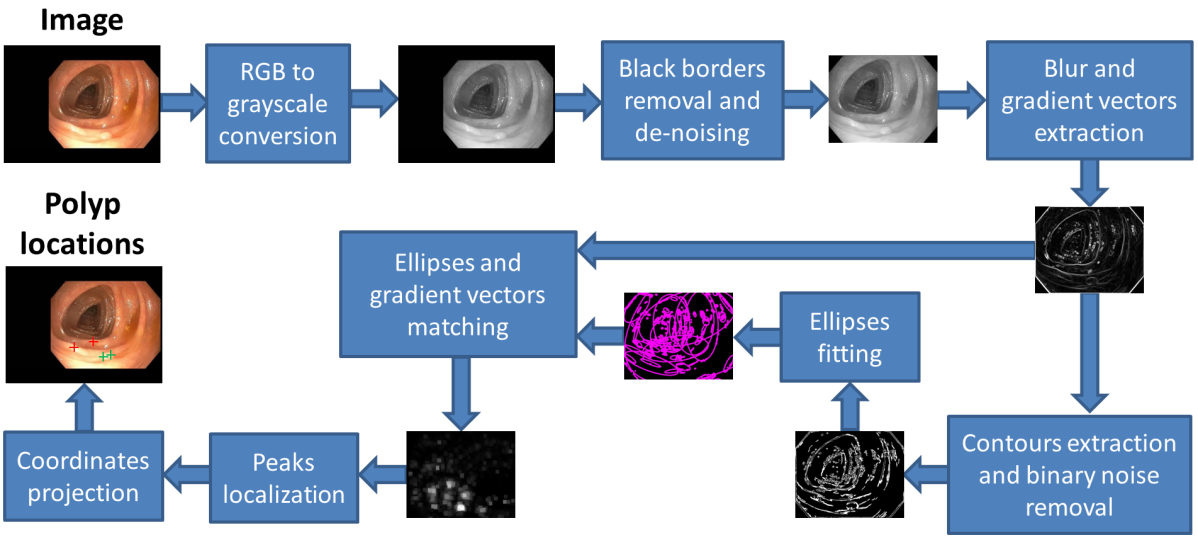


Figure 3.15: Detailed steps of the hand-crafted local-feature-based localization algorithm implementation

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 [115]. 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

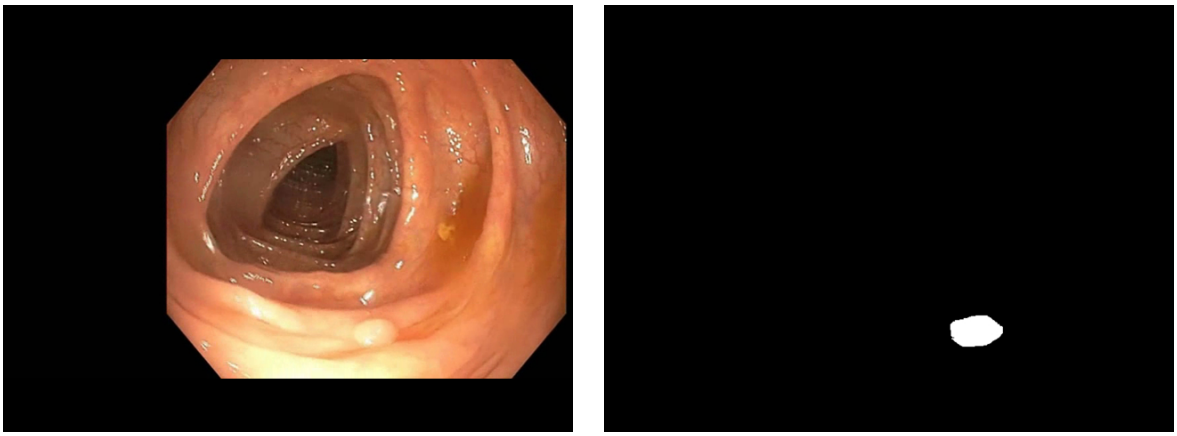
properties, like, eg., ulcer lesion and Z-line landmark. The next generation of complete localizers used in DeepEIR system are deep-learning-based region localizers. The idea of utilizing deep-learning-based methods for the localization tasks appeared in connection with the need to simplify support for 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] [127], 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 [130].

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 [130] 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, 127]. 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-

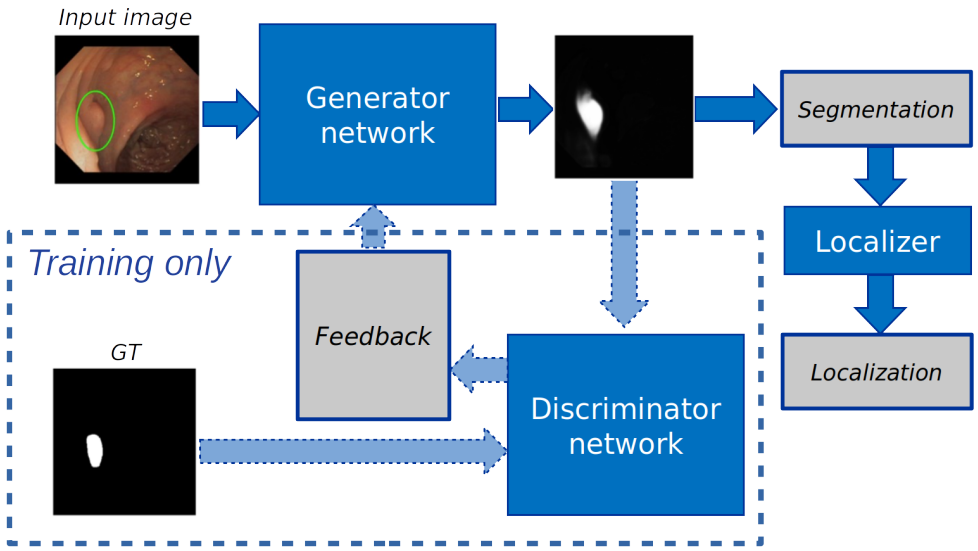
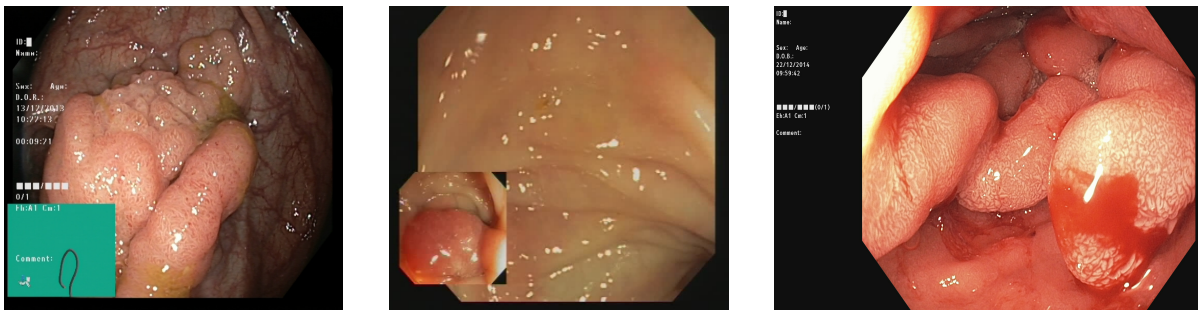


Figure 3.17: GAN-based segmentation and localization pipeline

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 [126] the good segmentation performance for the retinal images that have the visual properties comparable to the GI tract images. The V-GAN architecture [126] 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 $\pm 180°$, horizontal and vertical flipping of frames and image insensitivity alteration in the range of $\pm 40\%$. 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.

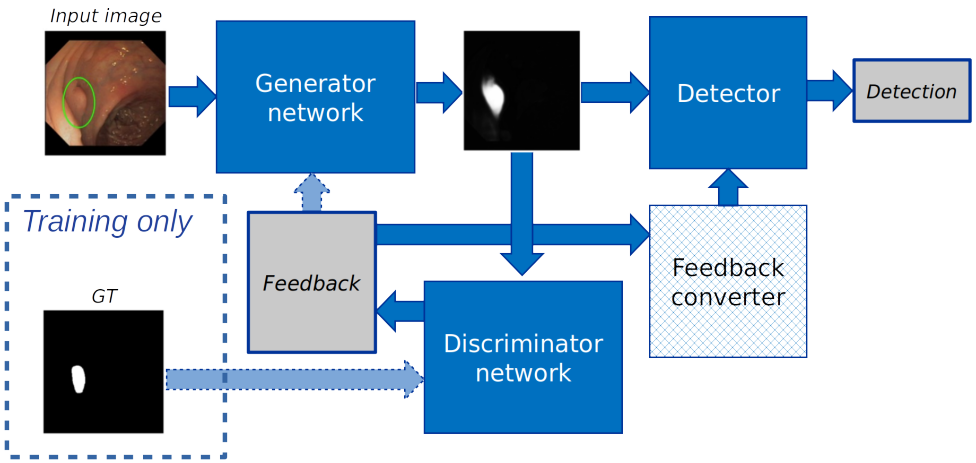


Figure 3.19: GAN-based detection-via-localization pipeline

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 [119] 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

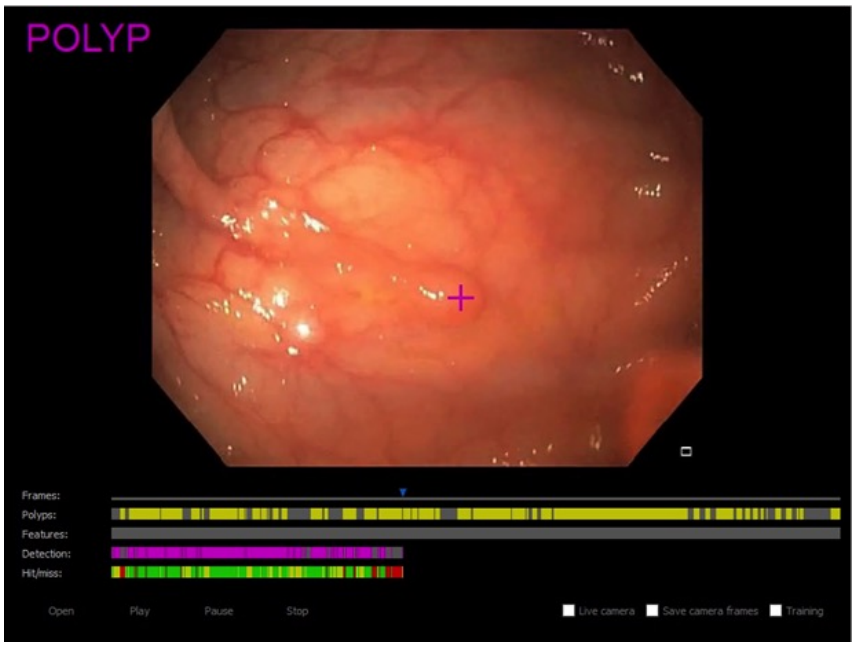


Figure 3.21: The visualization module for real-time polyp detection and spotting build upon our global-feature-based detection and hand-crafted local-feature-based polyp position finder approaches. It is able to process both recorded and live Full HD video stream from traditional colonoscope, highlight 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.

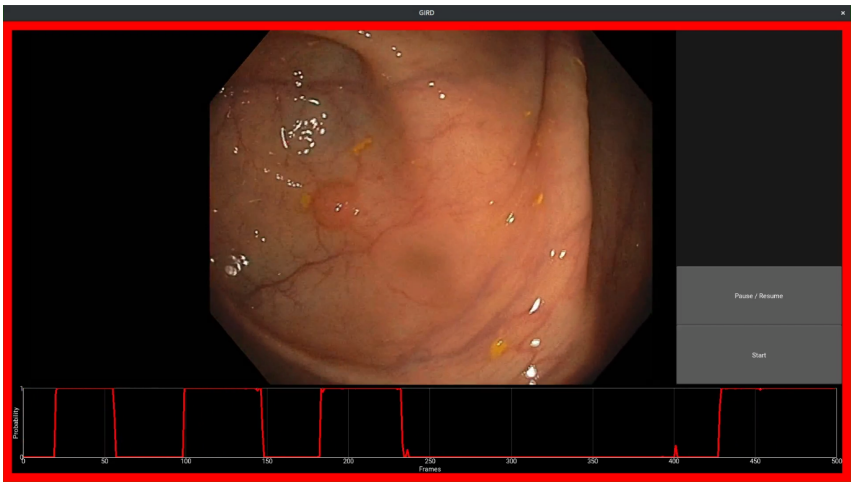


Figure 3.22: The visualization module for our deep-feature-based real-time polyp detection approaches. It is able to process Full HD live-captured video stream from traditional colonoscope and highlight frames containing detected lesions. The plot in the lower part of UI show the per-video-frame lesion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

inations 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].

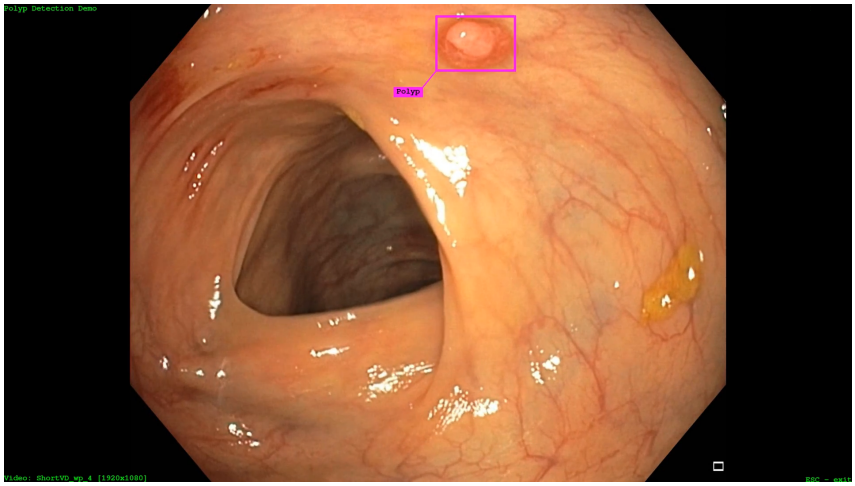


Figure 3.23: Near-to-real-time polyp detection and localization demo build upon our GAN-based detection and localization approach. The software processes recorded Full HD video stream from traditional colonoscope and highlights the exact polyp location in the particular frame. The marking is implemented 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].

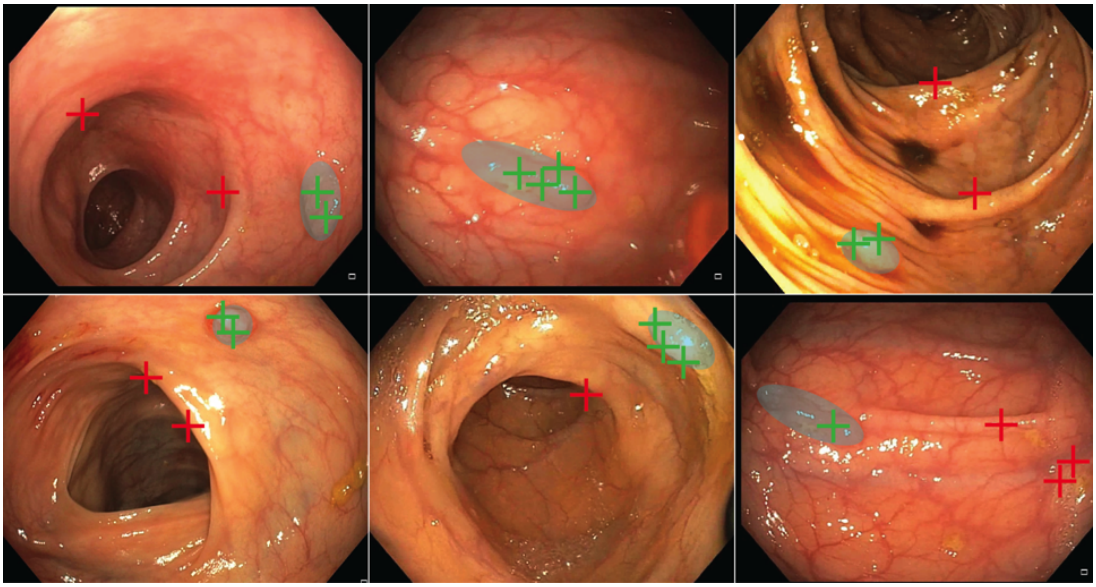


Figure 3.24: Polyp localization results generated by our first polyp localization and detection approach on the MICCAI 2015 dataset [25]. Light green ellipses depicts the polyp localization ground truth masks. Green and red crosses show the true positive and false positive polyp localization results, 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

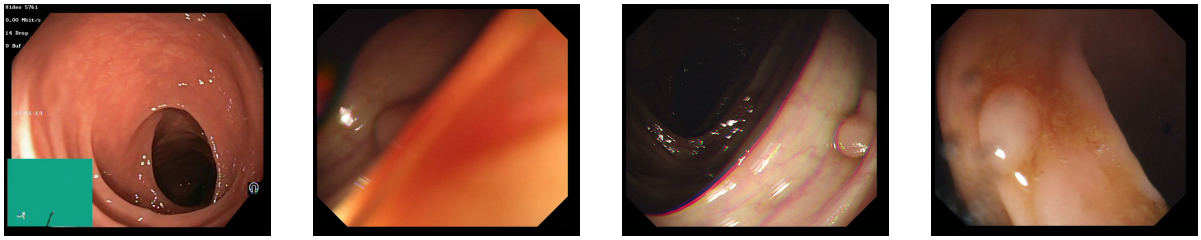
Table 3.8: This table depicts performance of the block-wise localization via detection approach for the CVC-12K dataset reported for different training data used.

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.

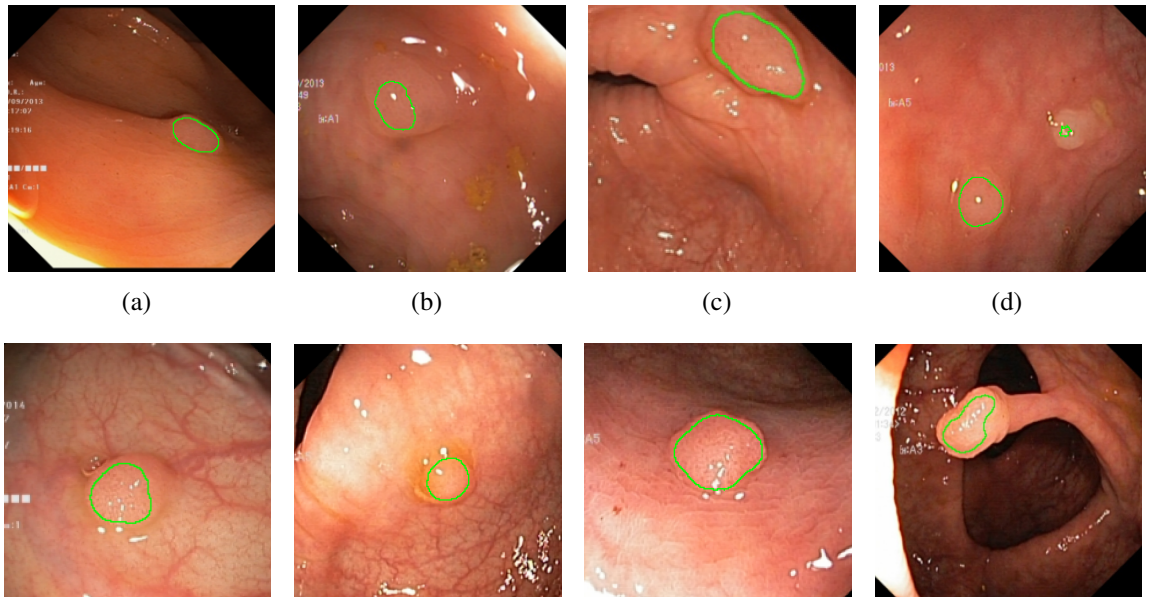


(a) Overlay image (b) Blurry frame (c) Colors shift (d) Lens contamination

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)

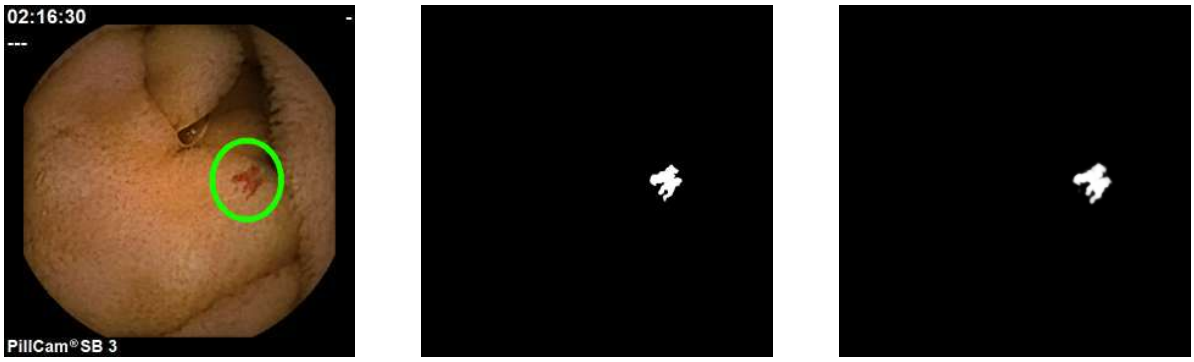
Figure 3.27: Examples of the detection and in-frame localization of the different polyps in the video frames captured by various vendors' traditional colonoscopy equipment. Green contour depicts the detected polyp and the localized main polyp body area.

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.

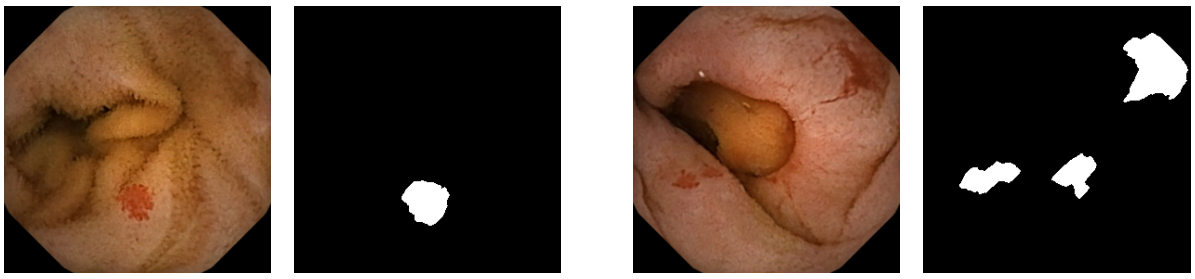
presented in Table 3.11. The detection outperforms significantly the 85% requirements. Both result sets are strong indicators that our GAN approach performs well for the tasks of angiectasia localization and detection.

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.

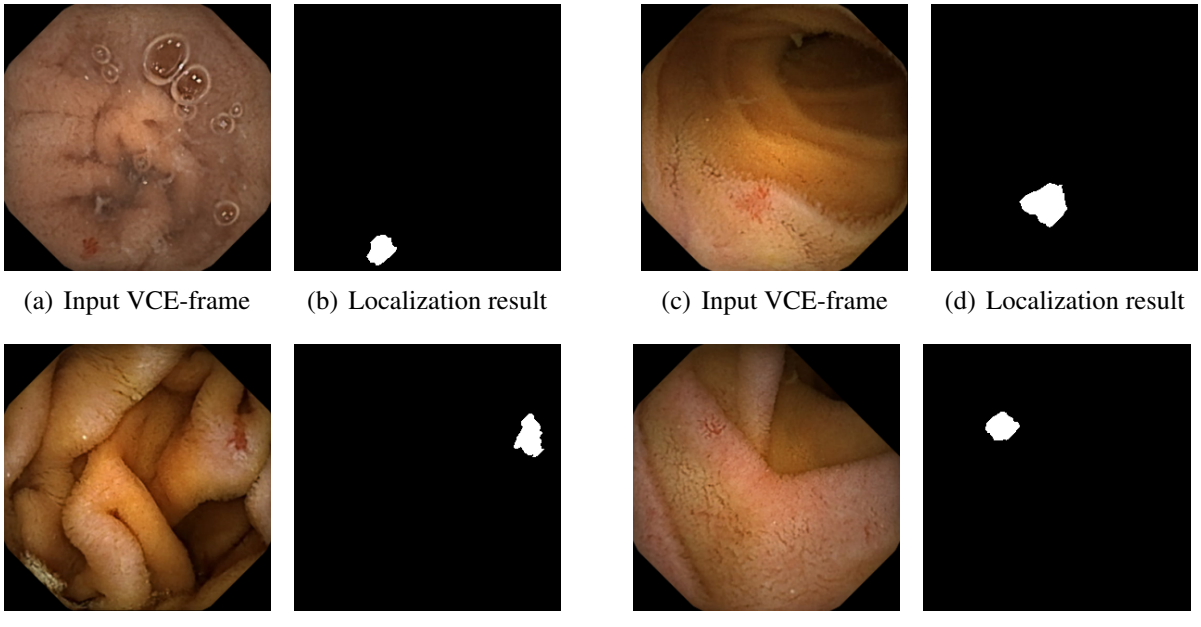
Approach	PREC	SENS	SPEC	ACC	F1	MCC	FPS
GF+LMT	0.695	0.680	0.680	0.680	0.674	0.375	80
DF ResNet50 CON+LMT	0.734	0.732	0.732	0.732	0.731	0.465	53
DF ResNet50 FEA+LMT	0.748	0.738	0.738	0.738	0.736	0.486	46
DF VGG19 CON+LMT	0.545	0.545	0.545	0.545	0.544	0.090	32
DF VGG19 FEA+LMT	0.525	0.525	0.525	0.525	0.525	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 [115] 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.

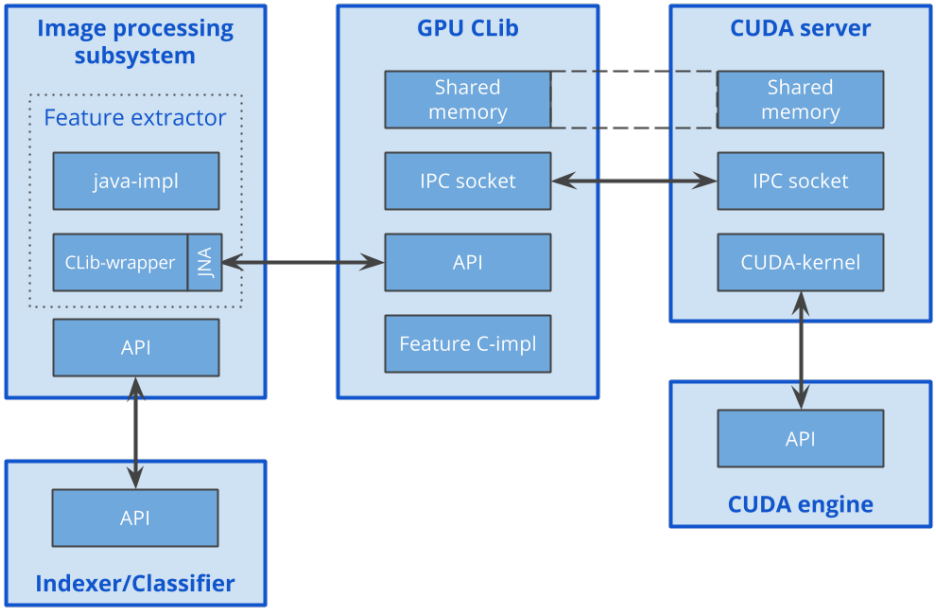


Figure 3.31: The main processing application consisting of the indexing and classification parts uses the GPU-accelerated image processing subsystem. This subsystem provides feature extraction and image filtering algorithms. The most compute-intensive procedures are executed on a stand-alone CUDA-enabled 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 prefiltering.

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

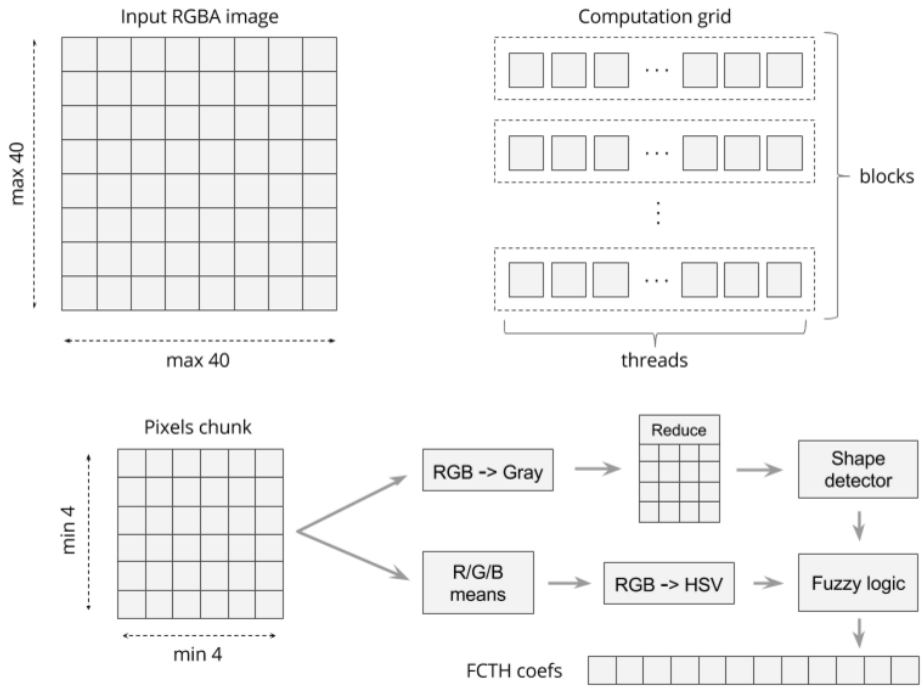


Figure 3.32: GPU-acceleration is used to extract various features from input frames. The figure shows an example of our FCTH feature implementation. The input frame is split into a number of non-overlapping blocks. Each of them is processed separately by two GPU-threads. The main processing steps include 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

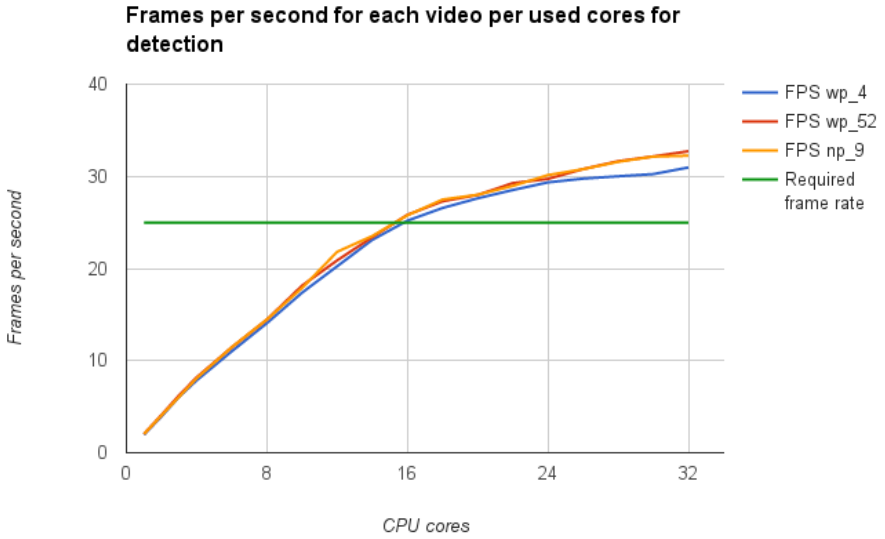


Figure 3.33: The detection performs efficiently and the required frame rate is reached with 12 GB of memory and 16 CPU cores used in parallel on cluster-based computation platform without utilizing heterogeneous architecture.

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

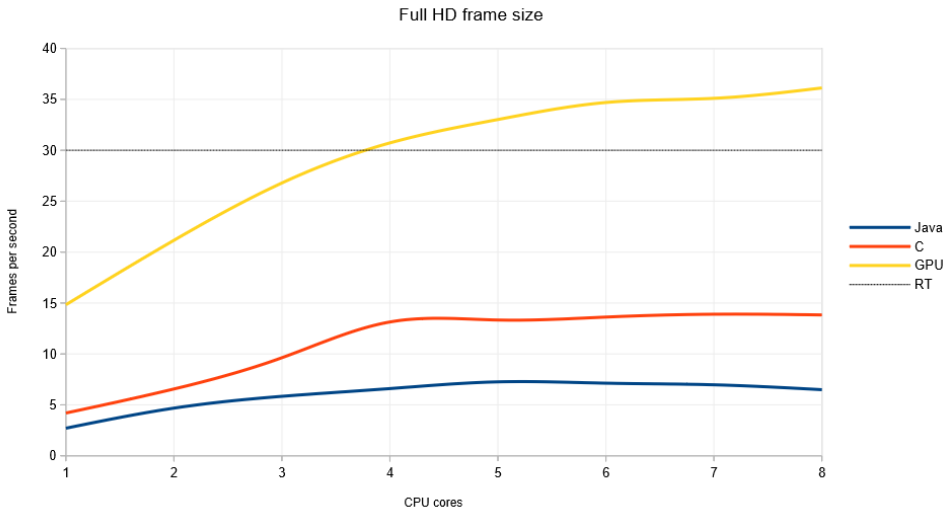


Figure 3.34: The improved GPU-enabled heterogeneous algorithm reaches real-time performance (RT line) with 30 frames per second for full HD (1920×1080) videos on a desktop PC using only 4 CPU cores and 5 Gb of memory. The maximum frame rate is around 36 FPS using 8 CPU cores. The Java and C implementations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

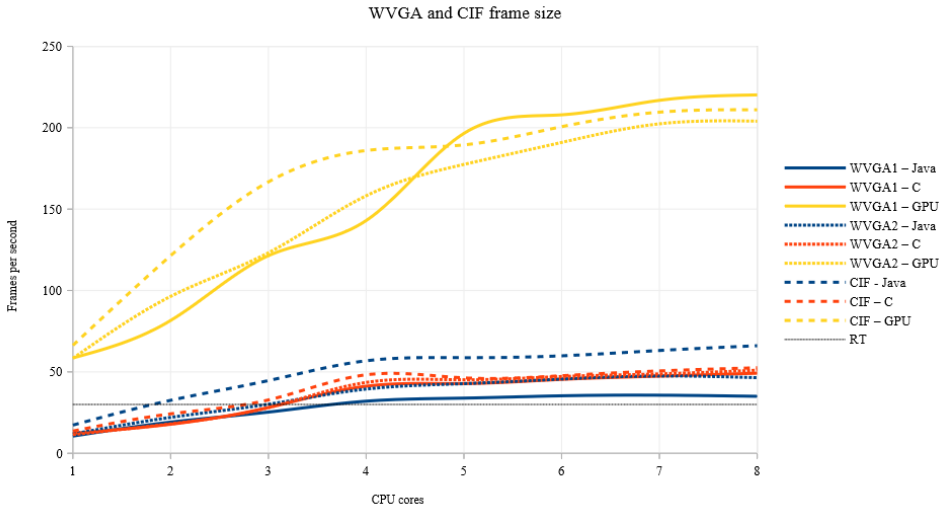


Figure 3.35: The smaller WVGA1 (856×480), WVGA2 (712×480) and CIF (384×288) videos can be processed by the improved GPU-enabled heterogeneous algorithm in real-time using only 1 CPU core. The maximum frame processing rate reaches more than 200 FPS. These results can be improved by putting all feature-related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

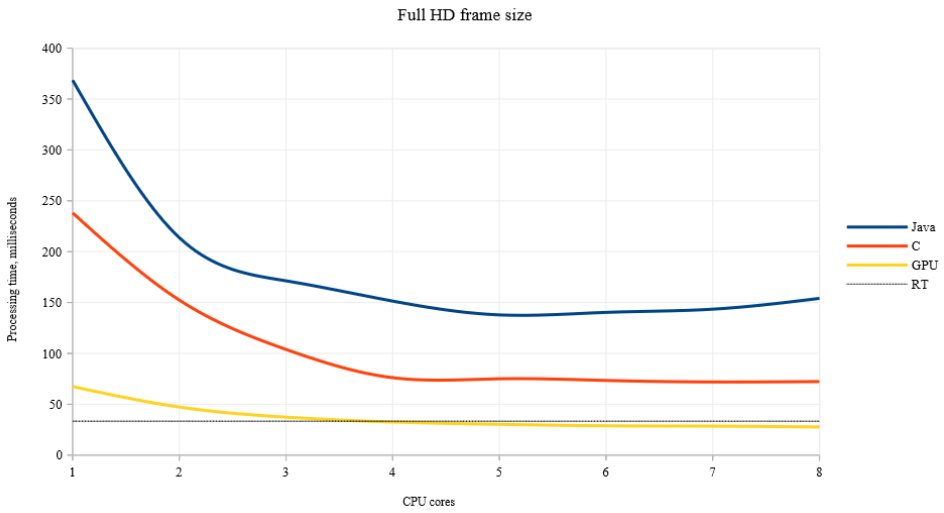


Figure 3.36: The processing time for the GPU-accelerated algorithm decreases slightly with increasing number of used CPU cores for a single full HD frame. This happens due to the CPU-parallel implementation of feature comparison and search algorithms which are not as compute intensive as feature extraction. 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,

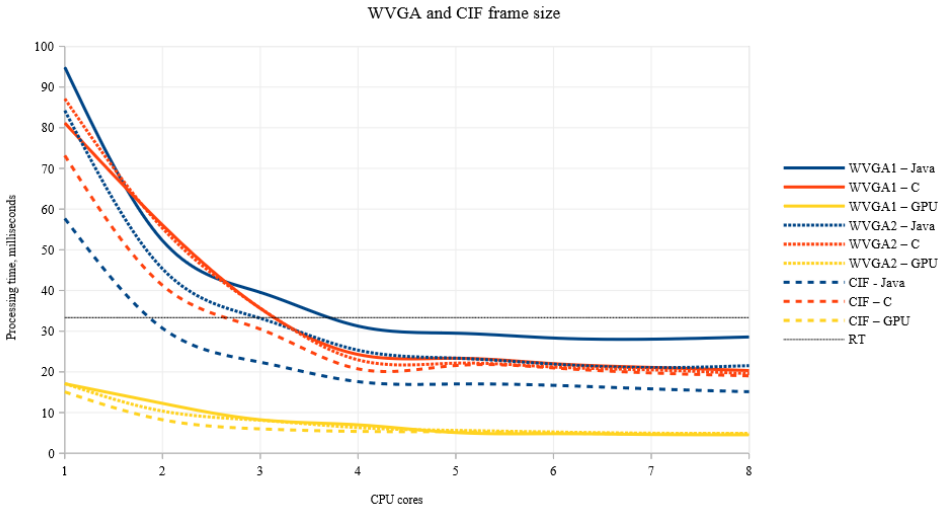


Figure 3.37: For the smaller frame sizes the GPU-accelerated algorithm results in a processing time far below the real-time margin. The minimum is reached with 5 milliseconds using 8 CPU cores. This is a prove for the high system performance and ability to be extended by additional features or to process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

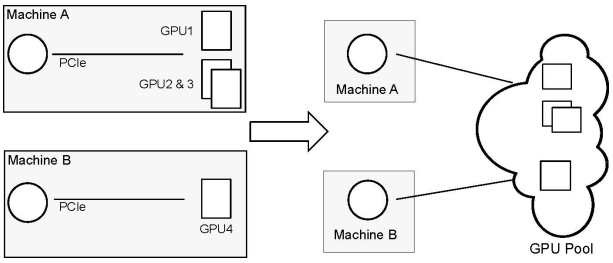


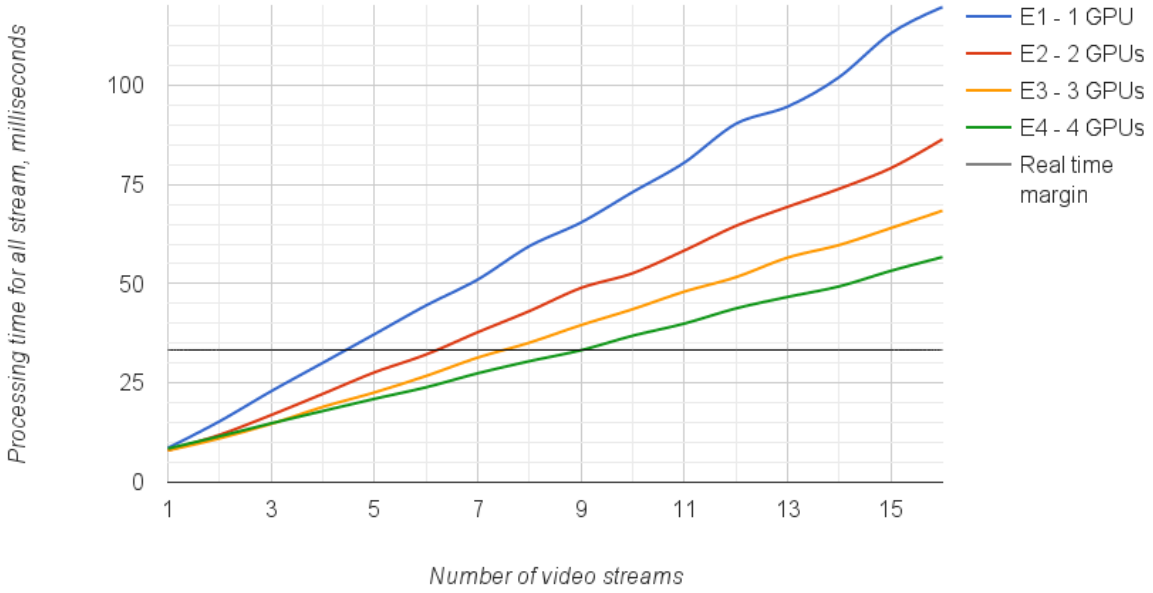
Figure 3.38: Pooling of devices attached in the PCIe network in the experimental setup.

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.

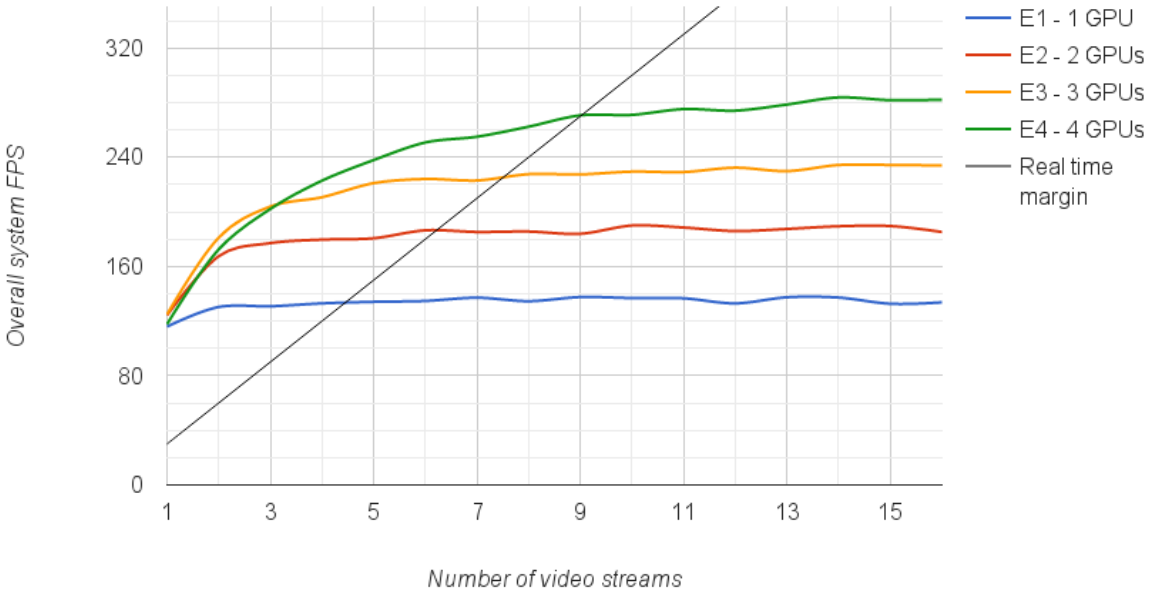
Device Lending is a concept where computers interconnected in a PCI Express [89] network can share devices. It provides transparent, low-latency cross-machine PCIe device sharing (see figure 3.38) without any need to implement application-specific distribution mechanisms or modify native device drivers. 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.

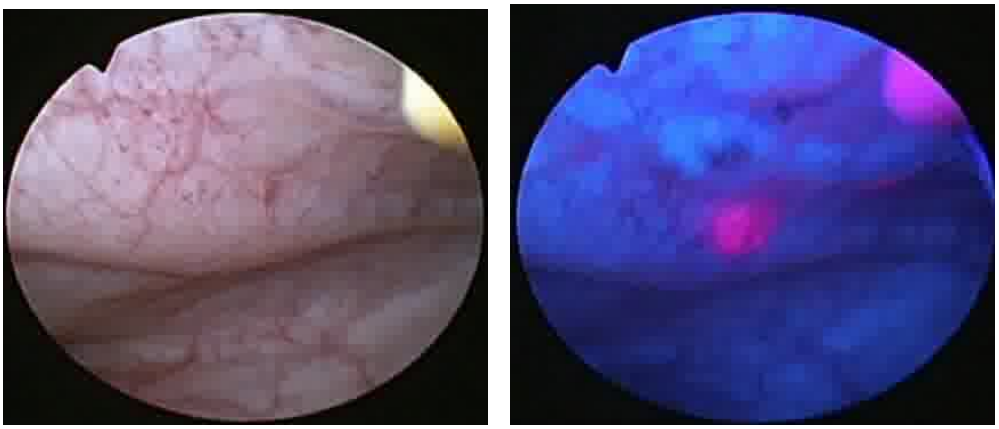
The plot in figure 3.39(b) shows the overall system performance. The maximum overall frames per second we reach when using 4 GPUs at the same time is 30 fps for 9 parallel full HD streams, which is equivalent to 270 fps for a single video stream. Further, this graph shows that the borrowed GPU does not 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 [123]. 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 [123]. 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.

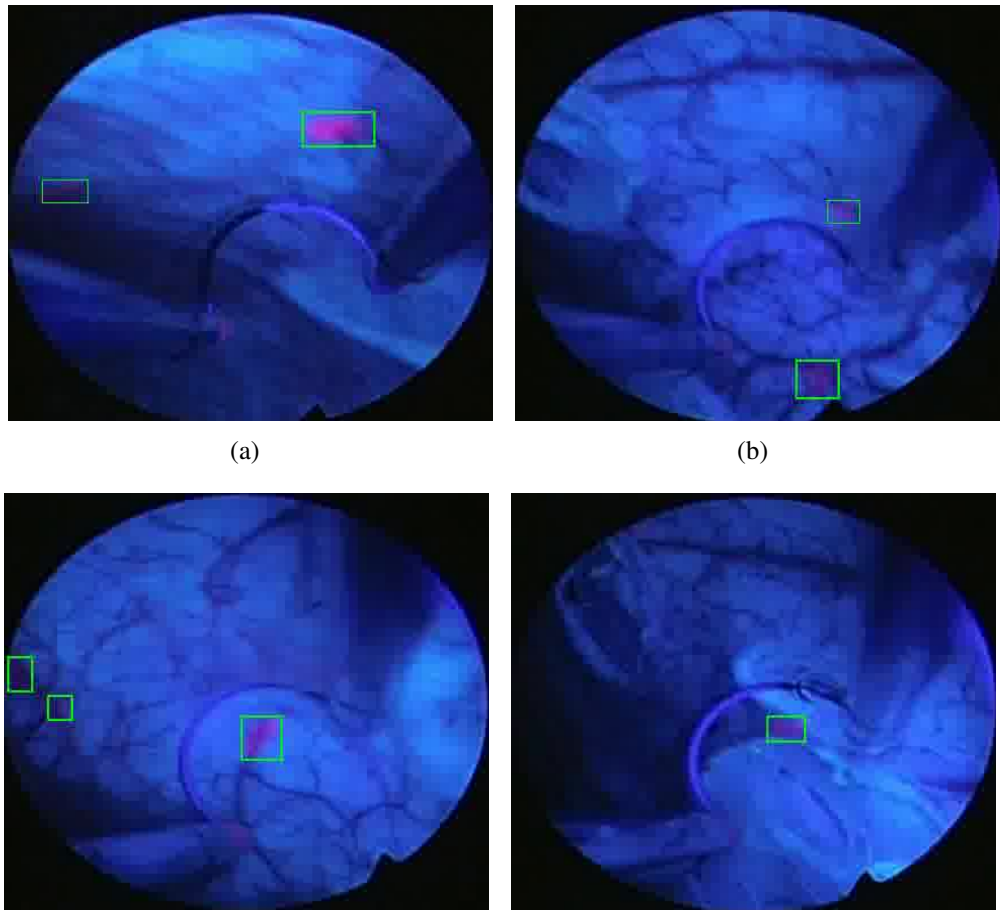
(c) BLC image depicts less visible tumor cells clusters partially be hidden by the interference with blood vessels.

(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 [146].

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,



(a) (b)

(c) (d)

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.

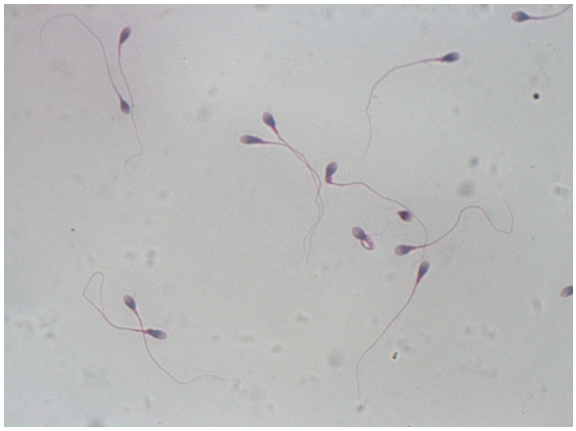
marking the areas showing the tumor cells. Using such a tiny training set allowed us to also test how our EIR system can deal with a new problem area with few annotated data samples, which is especially important for rare but still dangerous diseases.

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.

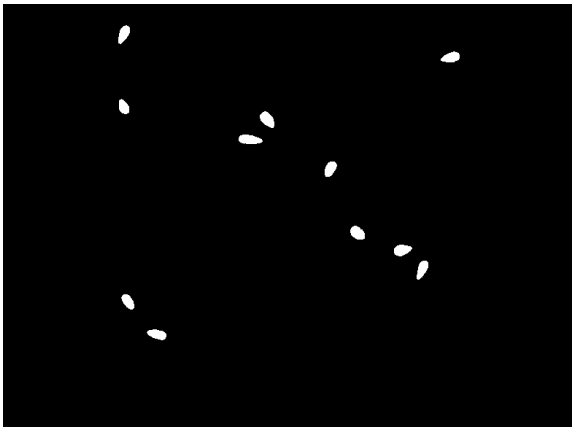
3.6.4.2 Spermatozoon Localization and Segmentation

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.

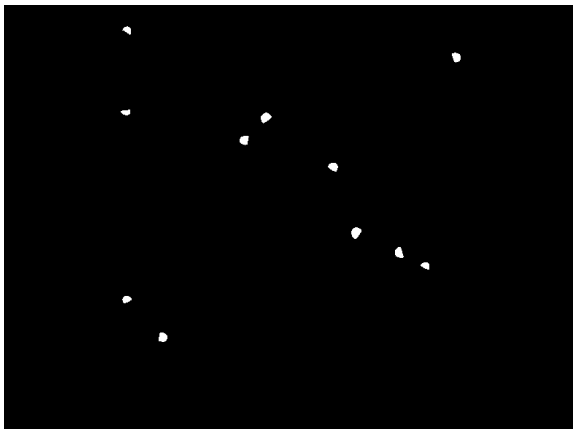
The crucial factor to the motility and morphology attribute measurement is the spermatozoon localization and morphological segmentation. For the morphology analysis, in the context of semen, doctors often examine the three parts that make up a spermatozoon. These include the



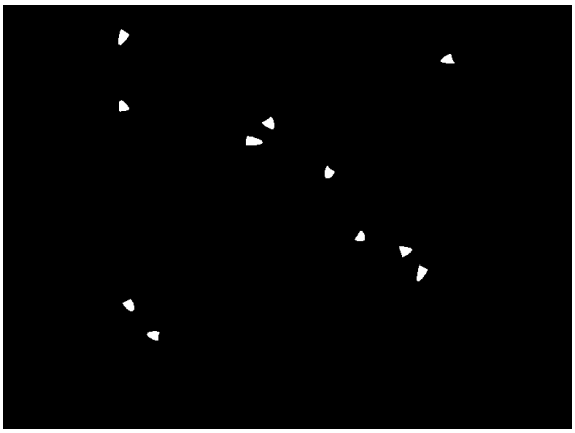
(a) Input microscopic image.



(b) Ground truth mask for heads.



(c) Ground truth mask for acrosomes.

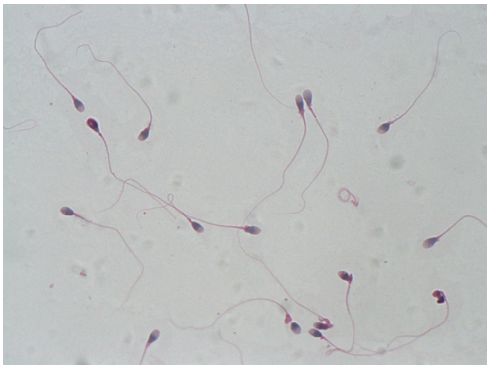


(d) Ground truth mask for nucleuses.

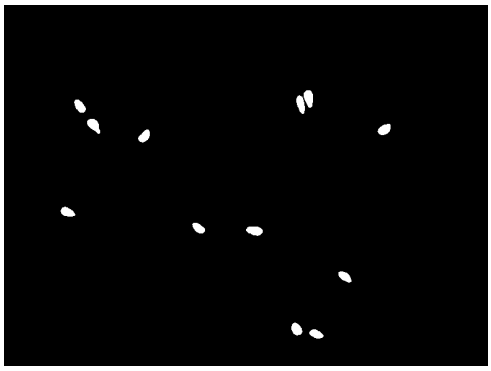
Figure 3.42: The example images of the spermatozoon localization and segmentation dataset used for the experimental evaluation of the EIR system with the different use-case study. First image (a) depicts the source microscopic image in RGB color space. Three other images (b-d) represent the ground truth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

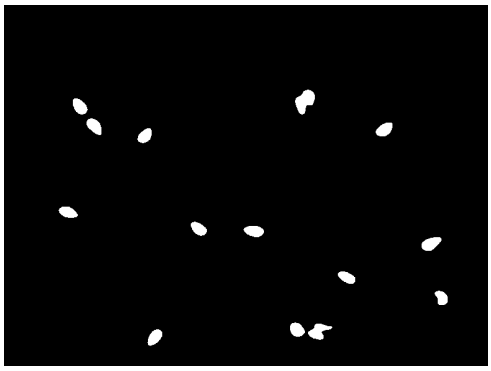
The dataset we used in the spermatozoon localization and segmentation experiment consists of 20 RGB frames recorded during a normal sperm microscopy procedure (see figure 3.42 for an example). Each microscopic frame comes along with three different ground truth masks for the different morphological parts 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



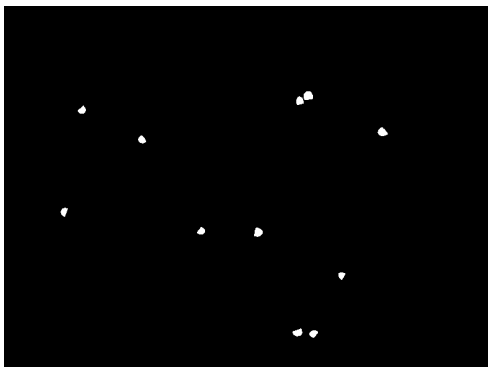
(a) Input microscopic image.



(b) Ground truth mask for heads.



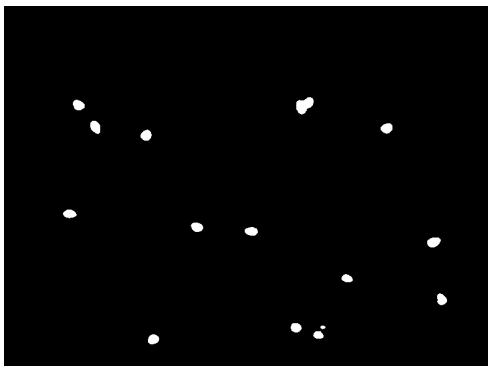
(c) Head segmentation results.



(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, 118].

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, 119].

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, 114] with new features and classification algorithms [115, 116]. 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 [116] 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, 118].

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 [128] 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 [117, 120]. 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 [116, 117].

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 [115, 116].

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, 116]. The whole system was tested by our collaborating medical doctors and was found promising and ready for clinical prototype development [91, 116]. 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, 121], 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 [115]. 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, 114, 116, 117, 118, 119, 120, 128] and side applications of its subsystems [13, 14, 15, 16, 55, 84, 113, 121], 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, 116, 117, 120].

  Using the ASU Mayo dataset [133], 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, 116, 117, 120].

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, 116, 120].

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, 116, 120].

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 [116]. 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, 120]. 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 [115].

We contributed to two open source projects: *LIRE*, in the field of content-based image retrieval [80], and *OpenVQ*, on video quality [125]. 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, 119].

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, 118]. 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 [128].

  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, 115, 116, 117, 120].

- **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 [120]. We developed an initial visualization approach that is able to visualize all outputs of the DeepEIR system [116], 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, 121] 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 [114].

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: 133

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: 139

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: 147

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: 153

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: 161

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: 169

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: 197

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: 209

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: 223

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: 231

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: 237

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: 273

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: 281

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 DeepEIR 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: 289

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: 295

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 postcolonoscopy 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, 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.

[115] 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.

[116] 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.

[117] 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.

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] 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.

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

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

[125] 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.

[126] 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.

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

[128] 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.

[129] 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.

[130] 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.

[131] C. Szegedy, V. Vanhoucke, S. Ioffe, et al. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015.

[132] 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.

[133] 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.

[134] 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.

[135] T. T. Tanimoto. elementary mathematical theory of classification and prediction. 1958.

[136] 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].

[137] 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].

[138] 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.

[139] 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.

[140] 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.

[141] 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.

[142] 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.

[143] 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.

[144] 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.

[145] 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.

[146] 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.

[147] 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].

[148] 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.

[149] 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 I

LIRE - Open Source Visual Information Retrieval

LIRE - Open Source Visual Information Retrieval

Mathias Lux
Klagenfurt University
Universitätsstraße 65-67
Klagenfurt, Austria
mlux@itec.aau.at

Michael Riegler, Pål Halvorsen, Konstantin Pogorelov
SIMULA Research
Oslo, Norway
michael@simula.no

Nektarios Anagnostopoulos
Klagenfurt University
Universitätsstraße 65-67
Klagenfurt, Austria
nek.anag@gmail.com

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 phones[1]. 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 *Lucene 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.

CCS Concepts

•Information systems → Multimedia information systems; Image search;

Keywords

Visual Information Retrieval; Search Engine

1. INTRODUCTION

Visual information retrieval and content based image retrieval have been around for years. In academia, it has been extensively reviewed (cp. [9]) and a lot of different approaches have been developed. However, early commercial software did not result in a broad application of visual information retrieval. Newer visual search engines took other approaches, like TinEye[2] with providing visual information retrieval technology as a service, or LegalZoom[3], which does a search for similar visual trademarks for the clients. Others focused on specific domains, like copyright infringement, medical retrieval, or near duplicate detection.

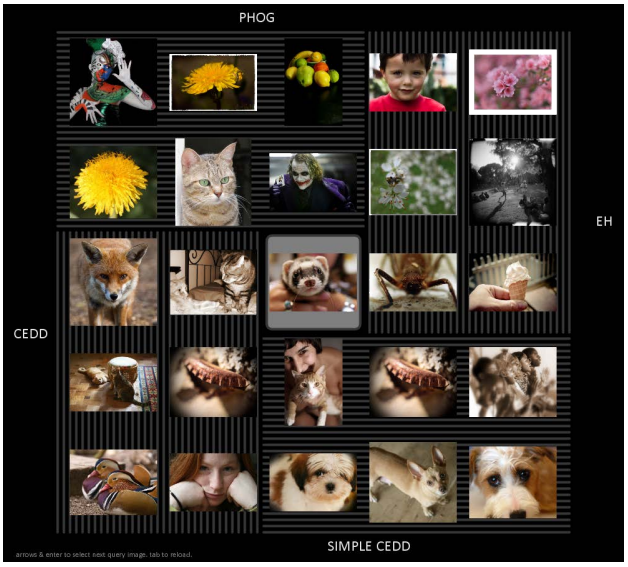


Figure 1: Sample application built on LIRE. The image in the center is the query, the first six results of four queries based on four different features, three global, one local one, are shown around the query.

However, nowadays, visual information retrieval builds on the academic achievements of successful research and a lot of different approaches, techniques and methods are available. Applied research then adapts the methods to new data and new domains. For this, it is crucial to have a common foundation that agrees upon algorithms and software implementations. Such a foundation can prevent developers and researchers alike from re-developing well-known approaches. A common, free and easy to access knowledge base is the main goal of LIRE.

LIRE provides the most common and well working approaches to content based image retrieval. Implemented as a Java library, it allows easy integration in existing software environments. LIRE builds on Lucene[4], which is a well-known and well maintained text search engine. Furthermore,

[1]http://goo.gl/nJz8gJ
[2]http://tineye.com

[3]https://www.legalzoom.com
[4]https://lucene.apache.org/

MMSys'16 May 10-13, 2016, Klagenfurt, Austria

ACM ISBN 978-1-4503-4297-1/16/05.
DOI: http://dx.doi.org/10.1145/2910017.2910630

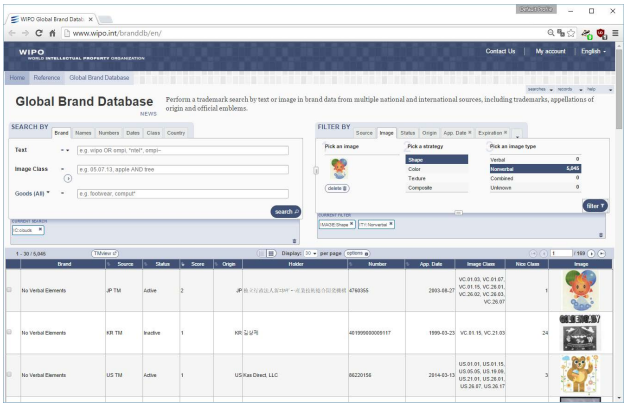


Figure 2: A screenshot of the UN WIPO Global Brand DB. The image filtering option is implemented using LIRE.

LIRE is the result of ongoing work of numerous contributors since February 2006. Since then, it is available as open source software under the GNU Public License. It has been hosted on sourceforge.net, Google Code and is currently maintained on Github[5]. Pre-compiled versions have been downloaded more than 51,000 times in 2015 alone. Major milestones were the release of the LIRE Solr Plugin in 2013 [16] and the version 1.0 beta release in 2015.

LIRE has been employed in academic research, teaching and real world scenarios alike. One major installation is at the UN headquarters in Geneve, Switzerland, running the visual trademark search at the World Intellectual Property Organization[6]. Fig. 2 shows a screen shot of the WIPO's Global Brand DB. There, a textual search for the term "clouds" is combined with a visual re-ranking based on a query image using PHOG [3]. Besides visual trademark search, LIRE has been employed for instance in asset management, copyright violation detection, and media monitoring. In the academic world, LIRE is used for feature extraction for classification, as base line for retrieval evaluation, for video search and summarization and as library providing image search for user interface and knowledge discovery projects.

2. LIRE

LIRE aims to be easy to use as well as easy to built new services on. If for instance new features are to be tested, developers and researchers only need to implement the feature interface including the serialization and extraction. Everything else then is done by LIRE, including parallel indexing, local feature aggregation, hashing, as well as approximate and linear search. This allows researchers and developers to focus on their features instead of having to implement the whole search engine.

LIRE supports multiple global and local features out of the box, to allow for easy comparison of new features to existing and well-known ones. Most notable global ones are CEDD [6] as well as the related features JCD [7] and FCTH [5], PHOG [3], the Auto Color Correlogram [11], Local Binary Patterns [18], CENTRIST [23], and the MPEG-7 features [4] Edge Histogram, Color Layout and Scalable Color.

[5]https://github.com/dermotte/lire
[6]http://www.wipo.int/branddb/en/

Local features are based on the OpenCV implementations of SIFT [15] and SURF [2]. For retrieval the bag of visual words approach [21] as well as VLAD aggregation of local features [14] are supported. In addition to that, LIRE fully implements the SIMPLE [12] approach to using global features on local image patches with configurable key point detectors.

For indexing, LIRE supports linear search as well as locality sensitive hashing [8] with a specific implementation of bit sampling. In addition to that, LIRE supports a permutation based approach called metric index [1], which adapts to image domains better than the hashing based approaches and employs inverted files for indexing [10].

3. PERFORMANCE

There are two main performance indicators for a image retrieval runtime: (i) performance on a single machine and (ii) scalability. For indexing, there are two main entry points. One is at the level of feature extraction, where indexing has to be handled by the users of LIRE. The more convenient approach is to use the parallel indexing routine provided by LIRE. It is configurable by supporting custom pre-processors, making use of multiple cores, and producing a Lucene index, which can easily be merged with indexes built with the same parameters. Thus, indexing is fully scalable.

For linear search, three optimizations are supported. These are, (i) memory cached search, where all image feature data is stored in memory, (ii) multi-core-search, where the search is run in parallel over index partitions, and (iii) DocValues based search using a mechanism of Lucene, where RAM and disk serialization are heavily optimized. With a GPU based approach, which is currently under development for indexing and searching video streams, indexes with up to one million images can be queried in 3ms for a resolution of $856x480$, and 18ms for images with a resolution of $1920x1080$. For more than a million images, LIRE provides approximate search techniques based on hashing [8] and permutation indexes [10]. Moreover, the index can be partitioned and search results can be merged to get more accurate results and at the same time increase speed [19].

Retrieval performance is shown in Table 3. The employed data sets are *SIMPLIcity* data set [22], the *UKBench Recognition Benchmark Images* data set [17], the *Uncompressed Colour Image Database* (UCID) [20], and the *INRIA Holidays* dataset [13]. While not being able to publish all possible feature and aggregation combinations, we aimed to give an overview on the performance. Retrieval features marked with a *(G)* in the Table 3 are global ones, i.e., Auto Color Correlogram, CEDD, Color Layout, Edge Histogram, JCD, Local Binary Patterns and Scalable Color. Global features marked with an *(SB)* are used on local image patches by employing the SIMPLE approach [12] with a bag of visual words aggregation. The number complementing the *SB* gives the number of visual words for this particular test. CVSIFT and CVSURF are the SIFT and SURF implementations from OpenCV, respectively. The *(B)* with the number indicates the use of the bag of visual words aggregation with the given number of visual words. *(V)* and *(SV)* denotes the use of the VLAD aggregation techniques for local and global features. In the latter case, the SIMPLE approach has been used to create local features first. The number of visual words is a lot smaller due to the VLAD aggregation.

	SIMPLICity [22]		UKBench [17]		UCID [20]		Holidays [13]	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
Auto Color Correlogram (SB, 128)	0.5380	0.7687	0.9082	0.3680	0.7752	0.2584	0.7914	0.2328
Auto Color Correlogram (G)	0.5099	0.7765	0.9253	0.3736	0.7488	0.2427	0.7986	0.2360
Auto Color Correlogram (SV, 16)	0.3920	0.7242	0.9009	0.3660	0.7513	0.2511	0.7602	0.2266
CEDD (SB, 2048)	0.5222	0.8030	0.8917	0.3596	0.7869	0.2611	0.7779	0.2284
CEDD (G)	0.5040	0.7410	0.8055	0.3324	0.6740	0.2229	0.7263	0.2114
CEDD (SV, 16)	0.4488	0.7333	0.8557	0.3504	0.7704	0.2542	0.7377	0.2154
CL (SB, 2048)	0.5211	0.7644	0.8399	0.3436	0.7079	0.2328	0.7385	0.2150
CL (G)	0.4506	0.6574	0.7035	0.2900	0.5675	0.1824	0.6480	0.1852
CL (SV, 64)	0.3747	0.6961	0.7844	0.3268	0.7068	0.2305	0.7060	0.2080
CVSIFT (B, 512)	0.3756	0.5620	0.6847	0.2808	0.6085	0.1954	0.6914	0.2016
CVSIFT (V, 64)	0.4489	0.6247	0.8047	0.3324	0.6933	0.2302	0.7581	0.2202
CVSURF (B, 2048)	0.3801	0.5555	0.6253	0.2644	0.5852	0.1885	0.6777	0.1954
CVSURF (V, 64)	0.4370	0.6111	0.6681	0.2900	0.6441	0.2145	0.7169	0.2092
Edge Histogram (G)	0.3454	0.5538	0.4832	0.2056	0.5019	0.1588	0.5551	0.1594
JCD (G)	0.5140	0.7498	0.8480	0.3464	0.6945	0.2279	0.7351	0.2162
Local Binary Patterns (G)	0.3699	0.6356	0.5302	0.2228	0.5325	0.1641	0.5575	0.1578
Scalable Color (G)	0.5222	0.7692	0.8990	0.3672	0.7116	0.2309	0.7454	0.2186

Table 1: Feature performance on four data sets. The X in (X) denotes: G for global, B for bag of visual words and V for VLAD aggregation. S for Simple, SB and SV denote bag of visual words or VLAD aggregation.

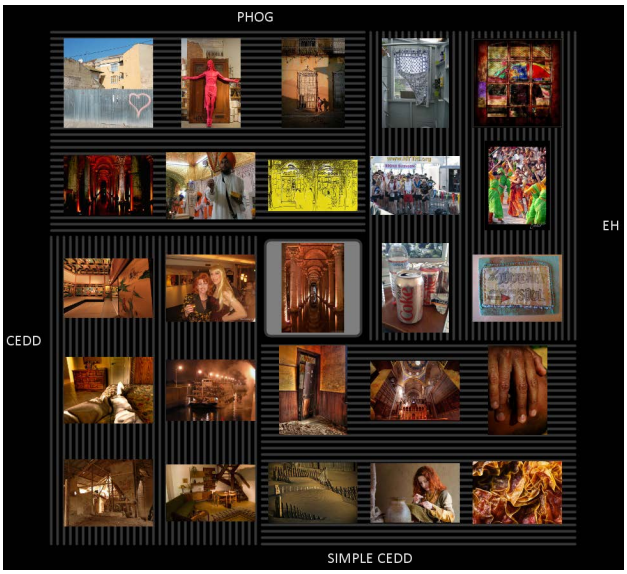


Figure 3: Sample application built on LIRE showing results for a different query image than Fig. 1.

4. DEMO

To show some of the aspects of LIRE, we present here a novel image retrieval and result browsing application. It utilizes the core strengths of LIRE: small footprint and minimal API, speed and accuracy. The difference to common image retrieval search engines is that it is a combination of 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 features. If users are for instance interested in similar colors and shapes, they can pre-select four different features that represent these attributes. 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. Fig. 1 and Fig. 3 show the desktop application. The query image is shown in the center, lines in the background of the results show the partitions. Users can navigate in the images and selecting an image results in a new search using the selected image as query. Therefore, users can browse the data set based on four different features. Artists and photographers for instance could find and browse images that share a either similar composition or color distribution at the same time. For example in Fig. 1 CEDD and SIMPLE CEDD give color based results with the latter providing different results as it is a localized version of CEDD, whereas PHOG and Edge Histogram (EH) based searches are returning images with similar composition. Fig. 3 shows the same composition of features for a different query image.

Moreover, we are testing the demo in a medical setting where it can help gastroenterologist (medical doctors specialized on the gastrointestinal tract of the human body) finding similar cases in their image databases. This is important since doctors are not likely to recall when and where a similar case happened, but they usually know if there was something similar in the past and how it approximately looked. The demo application is available for the desktop application written in *Processing 3* as well as for Android mobile phones and tablets.

Acknowledgments

We would like to thank at least some of the numerous people having contributed to LIRE: Anna-Maria Pasterk, Arthur Li, Arthur Pitman, Bart Van Bos, Bastian Hösch, Benjamin Sznajder, Berthold Daum, Carlos Perez Lara, Christian Penz, Christine Keim, Christoph Kofler, Chrysa Iakovidou, Dan Hanley, Daniel Pötzinger, Fabrizio Falchi, Franz Graf, Giuseppe Amato, Glenn MacStravic, James Charters, Ja-

nine Lachner, Katharina Tomanec, Konstantin Pogorelov, Lukas Esterle, Lukas Knoch, Manuel Oraze, Marco Bertini, Marian Kogler, Marko Keuschnig, Martha Larson, Michael Riegler, Nektarios Anagnostopoulos, Rodrigo Carvalho Rezende, Roman Divotkey, Roman Kern, Sandeep Gupta, and Savvas Chatzichristofis.

5. REFERENCES

[1] G. Amato and P. Savino. Approximate similarity search in metric spaces using inverted files. In *Proc. of InfoScale*, 2008.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. Springer, 2006.

[3] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 401–408, New York, NY, USA, 2007. ACM.

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

[5] S. Chatzichristofis, Y. S. Boutalis, et al. FCTH: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*, pages 191–196. IEEE, 2008.

[6] S. A. Chatzichristofis and Y. S. Boutalis. CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *Computer vision systems*, pages 312–322. Springer, 2008.

[7] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux. Selection of the proper compact composite descriptor for improving content based image retrieval. In *The Sixth IASTED International Conference on Signal Processing, Pattern Recognition and Applications SPPRA 2009*, 2009.

[8] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262. ACM, 2004.

[9] 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.

[10] C. Gennaro, G. Amato, P. Bolettieri, and P. Savino. An approach to content-based image retrieval based on the lucene search engine library. In *Research and Advanced Technology for Digital Libraries*, pages 55–66. Springer, 2010.

[11] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 762–768. IEEE, 1997.

[12] C. Iakovidou, N. Anagnostopoulos, Y. Boutalis, S. Chatzichristofis, et al. Searching images with mpeg-7 (& mpeg-7-like) powered localized descriptors: the simple answer to effective content based image retrieval. In *Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on*, pages 1–6. IEEE, 2014.

[13] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometry consistency for large scale image search-extended version. 2008.

[14] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010.

[15] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[16] M. Lux and G. Macstravic. The LIRE request handler: A Solr plug-in for large scale content based image retrieval. In C. Gurrin, F. Hopfgartner, W. Hurst, H. Johansen, H. Lee, and N. O'Connor, editors, *Proceedings of the 20th MultiMedia Modeling Conference (MMM 2014)*, volume 8326 of *Lecture Notes in Computer Science*, pages 374–377, Dublin, IE, Jan 2014. Springer International Publishing.

[17] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2161–2168. IEEE, 2006.

[18] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[19] 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 *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 397–406, New York, NY, USA, 2014. ACM.

[20] G. Schaefer and M. Stich. Ucid: an uncompressed color image database. In *Electronic Imaging 2004*, pages 472–480. International Society for Optics and Photonics, 2003.

[21] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[22] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(9):947–963, 2001.

[23] J. Wu and J. M. Rehg. CENTRIST: A visual descriptor for scene categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1489–1501, 2011.

Paper II

OpenSea - Open Search Based Classification Tool

OpenSea - Open Search Based Classification Tool

Konstantin Pogorelov
Simula Research Laboratory, Norway
University of Oslo, Norway

Zeno Albisser
Simula Research Laboratory, Norway
University of Oslo, Norway

Olga Ostroukhova
Research Institute of Multiprocessor Computation Systems n.a. A.V.Kalyaev, Russia

Mathias Lux
Klagenfurt University, Austria

Dag Johansen
UiT–The Arctic University of Norway

Pål Halvorsen
Simula Metropolitan Center for Digital Engineering, Norway
University of Oslo, Norway

Michael Riegler
Simula Metropolitan Center for Digital Engineering, Norway
University of Oslo, Norway

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.

CCS CONCEPTS

• **Information systems → Information retrieval**; • **Computing methodologies → Parallel algorithms**; **Search methodologies**; **Computer vision problems**; **Machine learning**;

KEYWORDS

Image, video, indexing, information retrieval, global features, machine learning, classification

ACM Reference format:
Konstantin Pogorelov, Zeno Albisser, Olga Ostroukhova, Mathias Lux, Dag Johansen, Pål Halvorsen, and Michael Riegler. 2018. OpenSea - Open Search Based Classification Tool. In *Proceedings of 9th ACM Multimedia Systems Conference, Amsterdam, Netherlands, June 12–15, 2018 (MMSys'18),* 6 pages.
https://doi.org/10.1145/3204949.3208128

Contact author's address: Konstantin Pogorelov, Simula Research Laboratory, Oslo, Norway, email: konstantin@simula.no .

MMSys'18, June 12–15, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5192-8/18/06…$0.00
https://doi.org/10.1145/3204949.3208128

1 INTRODUCTION

In the last years, multimedia data has become increasingly popular and important. More recently, big data is now a buzzword in the community related to the massive amount of multimedia data that becomes available because every user can create their own content and share it. However, also in other fields like medicine, the use of multimedia data, especially videos and images, has gained importance. This leads to a need for software and methods that make possible to search, categorize and classify this data efficiently based on content and not just based on metadata. Without such methods, the data available cannot be used efficiently. An example that leads to a lot of video data generation in the medical field is the use of camera pills (wireless video capsules), which traverse a patient's gastrointestinal (GI) tract. For a single patient, a camera pill collects between 4 and 12 hours of video material. Since medical experts are already overloaded, they do not have time to watch all the videos when the use of camera pills increases. Furthermore, batch processing of a huge amount of data costs resources and time, which are not always available. For cancer patients, it can be life-saving if their data is processed faster.

Therefore, we present an open-source system that provides a fast and easy way of classifying videos or images. It allows to easily create search-based classifiers that use global content features describing the image or frame as a whole. The OpenSea software contains a pipeline that allows the extraction of global features, creates indexes that are used as models for the classifier, classifies images and videos, and outputs the results in a format that can easily be used in a lot of various scenarios and applications by different users. Apart from that, OpenSea can also beat state-of-the-art methods such as convolutional neural networks (CNNs) in some use cases which makes it an easy to get and relatively hard to beat baseline for evaluation of approaches [6–8, 14]. It is also much faster compared to deep learning approaches [9]. Therefore, we believe that our tool is very useful for:

- Researchers who are not very familiar with classification, indexing or global features, but who want to use such methods.
- Researchers that develop state-of-the-art methods like deep learning architectures and want to benchmark their method with an easy to get and hard to beat baseline.
- Content managers and experts like medical doctors who want to use classification to learn from their data or understand their data better without going into technical details.

In addition, it can be useful for researchers who:

- Need a fast and reliable classification method that is easy to modify.
- Need just parts of the system like extraction of features, classification, etc.
- Want to evaluate their own classification methods based on comparison with our tool as an alternative to for example random baseline, etc.
- Work with big data and classification and segmentation problems.

In the remainder of the paper, we present the tool and how it can be used. We then present it in a medical use case to show its practical applicability.

2 THE SYSTEM

The tool that we have built consists of two separate parts, an *Indexer* and a *Classifier*, both parts are written in Java. To be able to extract image features from the content, we use the well known libraries OpenCV[1], Apache Lucene[2] and LIRE[3]. The *Indexer* can be used to create an index of images contained in a directory. The *Classifier* is able to read and process videos and images, and it uses feature similarities to perform a binary classification of frames in video or images in an index. The classification is done by identifying the most similar images in ground-truth indexes, which are provided as command line arguments.

2.1 Indexing and Training

The classifier uses indexes containing image descriptors of positive and negative examples as a model. Therefore, the classifier is trained by simply dividing negative and positive examples into the respective indexes. The Indexer accepts a list of directories as input in the command line. Each of the provided directories is then searched for image files, and an index of all the images contained in a directory is created. The index of all the images contained in a single directory is stored in a subdirectory called *index* in the form of Lucene-based indexes. The index can store multiple LIRE-feature-values per image, and the list of feature-values to store is provided in the command line. The supported features are all the features supported by LIRE [4] library. It is also built in a way that it can easily be extended in the case that one of the used libraries provides new features.

The usage of Lucene-based indexes has several advantages. These indexes are easy to compute and do not require a lot of storage space. Further, the indexes are optimized for search operations and can therefore be accessed efficiently. To increase the efficiency and the processing speed of our Indexer further, we have parallelized the indexing process. We create multiple threads that read image files from disk and calculate global image features concurrently. The results are then combined in a single index. The threads share the same list of files, but as the number of threads is fixed and known to each thread, we can split each video frame or image file statically. Every thread starts reading with an offset into one file and continues reading at offsets depending on its own thread ID. This allows us to implement reading without explicit locking of the input file list. Moreover, assuming that all images are of the same size, the workload is spread evenly across all threads. The actual number of threads used depends on the available processors reported by the Java Virtual Machine (JVM).

2.2 Classification of Video and Images

The classification of each video frame or image is based on the analysis of search results for a given query picture. The classification algorithm is a modified K-Nearest-Neighbor algorithm (k-NN). K-NN is a non-parametric algorithm, which means that the algorithm uses the rank of the values rather then the parameters of each frame. The frame classification is based on its k nearest neighbors by a majority decision. The classification algorithm used in the system differs in some points from the original k-NN algorithm. The first difference is that the algorithm is based on a ranked list of search results, which can be generated in real-time or pre-indexed for each query frame of the video. The second is that weighted values are used for generating a decision antithetical to the non-parametric behavior of the k-NN. The weights are based on the search result's ranked list. This part is designed in a way that it can easily be replaced with other different methods (for example visual page rank, etc.).

As mentioned before, the classification tool is implemented as a search for similar images in indexes that are generated off-line or on-the-fly, based on single or multiple image features. For every image in the input index or video, it searches the provided classifier indexes and finds the images with the most similar image features, whereas similarity is determined based on the low level features and their associated distance (in this case Tanimoto distance). Based on the class of the similar images retrieved from the index, the input image is classified. The result for every single image feature, as well as the result of late fusion for all the selected image features is displayed on-screen. Late fusion means that each feature has an own classification step that is combined with other classifiers' output for the final result. When classifying previously indexed images, an HTML page is created with a visual representation of all the classified images. When classifying a video sequence, the results are stored to a file in JSON format instead. The classification tool also determines the performance of the classification and calculates several evaluation scores such as *precision, recall, weighted f1-score, etc.*. For this to work, the input data must be labeled correctly before it is classified. This can either be done by prefixing the filenames of the files in the test index with 'p' or 'n' for positive and negative samples, respectively, or by supplying separate test indexes with the command line options for the input data.

[1] http://opencv.org/ [last visited, Feb. 10, 2018]
[2] http://lucene.apache.org/ [last visited, Feb. 10, 2018]
[3] http://www.lire-project.net/ [last visited, Feb. 10, 2018]

3 INSTALLATION AND LICENSE

OpenSea is licensed under the terms of the GNU General Public License (GPL) version 3, as published by the Free Software Foundation. OpenSea depends on LIRE, which is licensed under GPL version 2, and OpenCV, which is licensed under a BSD license.

We have tested our software on *Linux*, *Mac OS X* and *Windows*. For simplicity, we provide installation instructions for *Ubuntu Linux*. All the required files from stable LIRE version 0.9.5 and Apache Lucene distribution are already included into the OpenSea distribution. The following installation and build instructions were tested with *Ubuntu 16.04*:

- Download and install the Java SE Development Kit 8 from http://www.oracle.com.
- Make sure to have the directory containing the java compiler in your PATH environment variable.
- Install *OpenCV-Java* and *Apache ant*:
  `sudo apt-get install libopencv2.4-java ant`
- Clone the OpenSea repository:
  ```
  git clone \
      https://github.com/acmmmsys/2018-OpenSea
  ```
- Build OpenSea using *ant* as command line arguments.
  `ant dist`
  Once building finished, you should find the two files *classifier.jar* and *indexer.jar* in the subdirectory *dist*.
- To make sure the *OpenCV-Java* native libraries are found at runtime, it is further necessary to add the path to *libopencv_java249.so* to *LD_LIBRARY_PATH*.
  `export LD_LIBRARY_PATH=/usr/lib/jni`

If another versions of LIRE is required, the following additional steps are required:

- Download *Lire* from http://www.lire-project.net/.
- Unzip *Lire* to a directory of your choice. We will refer to this location as *Lire directory*.
- Make sure your *LIRE directory* contains the file *lire.jar*.
- Build OpenSea using *ant*, passing your *Lire directory* and the corresponding *OpenCV-Java directory* as command line arguments. The *OpenCV-Java directory* is where your java bindings for *OpenCV* were installed (used by both LIRE and OpenSea). It must contain the file *opencv-249.jar*, or any later version.
  ```
  ant -Dlire=/home/me/Lire \
      -Dopencv=/usr/share/OpenCV/java dist
  ```
  Once building finished, you should find the two files *classifier.jar* and *indexer.jar* in the subdirectory *dist*.

4 USAGE INSTRUCTIONS AND EXAMPLE

To show how to use OpenSea, we provide the usage instructions and a few command line examples.

4.1 Indexing

The indexer can be started as follows:

```
java \
        -jar [/path/to/jar/file/]indexer.jar
        [-f feature]
        /dir/with/images
        [/dir/with/more/images]
```

Indexer support multiple features set by *-f* command line argument as well as multiple directories with images or frames extracted from video.

Usage example:

```
java \
        -jar indexer.jar -f JCD -f FCTH \
        /home/user/dataset/train/pos \
        /home/user/dataset/train/neg \
        /home/user/dataset/test
```

This creates the two indexes containing the global image features *Joint Composite Descriptor* (JCD) and *Fuzzy Color and Texture Histogram* (FCTH) of the images in the */home/user/dataset/train/pos*, */home/user/dataset/train/neg /home/user/dataset/test* directories and stores the indexes in */home/user/dataset/train/pos/index*, */home/user/dataset/train/neg/index* and */home/user/dataset/test/index* directories respectively. If the index target directories contain any previously extracted features they will be replaced.

4.2 Classification

The classifier can be started as follows:

```
java \
        [-Djava.library.path=/path/to/opencv/for/
            java]
        -jar [/path/to/jar/file/]classifier.jar
        [-f feature]
        [-c /dir/with/training/index]
        [-p /dir/with/training/positive/index]
        [-n /dir/with/training/negative/index]
        [-i /dir/with/test/index]
        [-P /dir/with/test/positive/index]
        [-N /dir/with/test/negative/index]
        [-v /path/to/video/file]
```

Classifier support multiple features set by *-f* command line argument. Training and test datasets are expected to be supplied in the indexes previously extracted by Indexed. Indexes can be either joint or separated sets of positive and negative samples. For the joined sets file names must start with 'p' for positive sample and with 'n' for negative samples, with corresponding *-c* and *-i* command line arguments for training and test sets respectively. For the separated sets positive and negative samples must be provided in the separate indexes, with corresponding pairs *-p*, *-n* and *-P*, *-N* of command line arguments for training and test sets respectively. Classifying of video frames is implemented via *-v* command line arguments which is mutually exclusive with test set arguments.

Usage example:

```
java \
        -Djava.library.path=/usr/lib/jni \
        -jar classifier.jar \
        -f JCD -f FCTH -f Tamura \
    -p /home/user/dataset/train/pos/index \
    -n /home/user/dataset/train/neg/index \
    -i /home/user/dataset/test/index
```

This example shows how to classify images from the index */home/user/dataset/test/index* using the image features *JCD* and *FCTH*, by finding the most similar images among the positive samples from */home/user/dataset/train/pos/index* and the negative samples from

/home/user/dataset/train/neg/index. For the calculation of the evaluation metrics, it is required that the images indexed in */home/user/-dataset/test/index* have names starting with 'p' or 'n' for positive or negative samples, respectively. This generates visual classification output in HTML format. Example of generated HTML is depicted in figure 1.

```
java \
        -Djava.library.path=/usr/lib/jni \
        -jar classifier.jar \
        -f JCD \
    -p /home/user/dataset/train/pos/index \
    -n /home/user/dataset/train/neg/index \
    -P /home/user/dataset/test/pos/index \
    -N /home/user/dataset/test/neg/index \
    -f JCD
```

The second example uses samples from the positive index */home-/user/dataset/test/pos/index* and negative samples from the negative index */home/user/dataset/test/neg/index*, which are classified using the image feature *JCD*. The previously known classification is only used for evaluating the results of the classifier.

```
java \
        -Djava.library.path=/usr/lib/jni \
        -jar classifier.jar \
        -f JCD \
    -p /home/user/dataset/train/pos/index \
    -n /home/user/dataset/train/neg/index \
    -v /home/user/dataset/testvideo.avi
```

In our last example, a video file is supplied as input to the classifier. All video frames of this input video */home/user/dataset/testvideo.avi* are classified by searching the most similar images among the positive samples from */home/user/dataset/train/pos/index* and the negative samples from */home/user/dataset/train/neg/index* using the global image feature *JCD*. In addition to the on-screen output (see figure 2 for an example), a JSON file is generated, which contains a list of all the positive frames and a list of all the negative ones.

To process videos in real-time, we have also parallelized the classifier. Again, the number of threads created depends on the number of processors reported by the JVM. Each thread holds a separate instance of the classifier indexes, but all threads share the same queue for the input data to be classified. Therefore, every image or video frame is only loaded once, it is then processed by a single thread, and the result is written to a shared data structure. This allows for all threads to operate independently, with only two critical sections, i.e., one for dequeuing the next input image and one for writing to the shared result data structure. When processing a video as input data, an additional thread is created for reading the video from a file and filling the input frame queue. The Classifier tool further provides different options for weighting the count or distance score of similarity results. The different weighting methods can be chosen by adding the flag $-m$ followed by the rank method that should be applied to the command. As default mode, no weight is set and the classifier uses only the count per class. At the moment, we support 3 additional weighting methods: (i) weighted by rank position, i.e., the weight is computing from the position in the returned ranked list, (ii) weighted by distance, which uses the Tanimoto distance from the search as weight and (iii) weighted by average distance, which uses the average distance of all returned

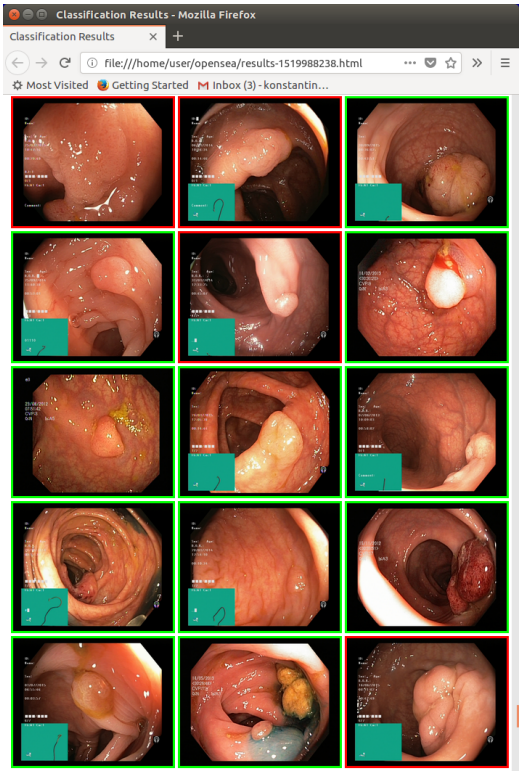


Figure 1: Example of a generates visual classification output in HTML format. Images with green borders correspond to true positive and true negative samples. Images with red borders correspond to false positive and false negative samples.

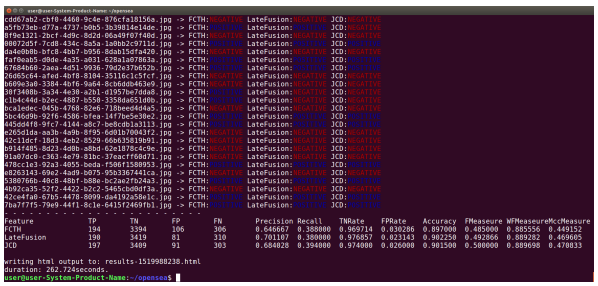


Figure 2: Example of a classifier on-screen output. The output contains classification results for each used feature and features' late fusion, as well as the corresponding performance metrics.

documents in the ranked list instead of the number of documents to calculate the weight. Moreover, various different combinations of global image features can be evaluated separately or combined in late fusion. This makes the tool ideal for experimenting with different approaches and finding an optimal set of features to use for a specific use case.

4.3 Metrics

In our performance evaluation, several specific metrics are implemented and can be calculated for the test data T. All metrics are calculated based on true positives (tp), false positives (fp), true negatives (tn), false negatives (fn) values per class $c \in T$. The most important metrics are precision, recall, $F1$ score and Weighted $F1$ score. A common and often used metric to calculate the quality of a classifier that considers both measures, *precision* and *recall*, is the $F1$ score. It is the harmonic mean (mean value of a number of values) of *precision* and *recall*. One problem with the standard $F1$ score is that a low value is not always an indicator of a badly performing classifier or retrieval system if the classes in the test dataset are not normally distributed [13]. To solve this problem, the weighted $F1$ score (WF1) can be used. This score takes both, the negative and the positive class results, into account and calculates a more accurate and robust measure. $WF1$ score is known to be more reliable to evaluate the performance of a classifier or retrieval system than the standard $F1$ score. Apart from $F1$ and $WF1$, the tool also provides true negative rate, false positive rate, accuracy and the Matthews correlation coefficient [5].

All of these metrics are suitable for showing the performance of binary (two classes) and multi-class classifiers (more than two classes) and should be a valuable set of instruments for users of the tool to evaluate their classifiers.

5 USE CASE

To show how the system works, we performed two experiments using two different pairs of training and test sets. For the first experiment, we used the ASU-Mayo Clinic polyp database [2]. It is at the moment one of the largest publicly available dataset of colonoscopy videos. The dataset comes with a ground truth that indicates if a frame of a video contains a polyp in the colon or not. The dataset consist of 20 videos. 10 videos do not contain polyps at all, and 10 of them contain polyps in the whole video or parts of it.

First, we split the dataset into test and training sets. The test set contained two separate videos that are not used in the training dataset. To measure the performance, we used the well known metrics precision and recall. All the tests were conducted without a weighting method (default mode). In this first test, we achieved a precision of 0.903, a recall of 0.919. For these results, we used a fusion of the features JCD and OpponentHistogram, which we found to perform best in small tests before [15]. The number of visual neighbors (size of the rank list returned by the search part of the classifier) was 71. The majority class baseline (all negative) is $0,683$ for precision, $0,683$ for recall.

To evaluate the robustness of the classifier, and to check if the good results were not just overfitting, we decided to perform a leave-one-out-cross-validation (LOOC) with all 20 videos of the dataset. In LOOC, all videos of the dataset are used to train the model expect for one that is used as the test example. This is repeated, so that all the sample videos are excluded once. To be able to recreate the experiments and test the software, we added the indexes to the official repository. We used the same features and number of visual neighbours as in the test before. For LOOC, the average precision is $0,895$, the average recall is 0.903. In comparison to the LOOC for the majority class baseline (all negative) which reaches a precision of 0.636, recall of 0.636. It is important to point out that we choose the class with the highest number for the majority vote baseline against the common practice to decide for the positive one. This makes it harder to outperform the baseline, but it also shows the real performance of the classifier. The results shows that our system performs well in cross validation and that it is robust and not overfitted for the dataset. We also want to point out that the classification time is very low. For a single frame, the time is around 30 milliseconds. To be able to do it in real time for videos with 30 frames per second, $33,3$ milliseconds is the deadline. In the best case, if we use a single feature, we can even get a classification time of around 10 milliseconds. The parallelization is not yet optimized, and we have some ideas that can make the system even faster, but this is out of scope for this paper.

Table 1: A performance comparison of deep-learning and global features based GI findings detection approaches [9]

	Global-features-based EIR	*Deep-EIR*
Detection Type	polyps / 30 features	abnormalities / neural network
Recall (Sensitivity)	98.50%	87.20%
Precision	93.88%	87.20%
Specificity	72.49%	97.40%
Accuracy	87.70%	97.50%
FPS	**300**	**30**

For the second experiment, we use combinations of four different, publicly available datasets, namely CVC-356 [2], CVC-612 [1], Kvasir [8] and parts of Nerthus [7]. The CVC-356 and CVC-612 datasets consist of 356 and 612 video frames, respectively. Each frame that contains a polyp comes with pixel-wise annotations in the CVC-356 and CVC-612 datasets. They both are used for training only in our polyp detection experiments. The fames from those datasets were renamed adding 'n' or 'p' prefix to reflect actual polyp presence in frames according to the existing pixel-wise annotations. For the testing we used Kvasir and Nerthus. For the Kvasir dataset, we included all classes except for the dyed classes (in a real world scenario something dyed is already detected by the doctor) leading to a dataset containing 1,000 frames with polyps, 5,000 without. We also added the 1,350 of class three frames with normal mucosa from the Nerthus dataset.

For this experiment we performed training and polyp detection using the described sets with two different detection approaches: the proposed OpenSea system and a deep-learning based abnormality detection approach [9]. The comparison of performance and data processing speed is depicted in table 1. As one can see, the OpenSea (global-features-based EIR) approach can perform as good in terms of detection performance as deep-learning based, but OpenSea system perform ten times faster in terms of processing speed. This results showing a promising nature of global features and their ability to perform fast and efficiently even across the different datasets.

The problem of polyp detection in GI videos is one of the most important problems in modern medical endoscopic imaging analysis [6]. Our efforts in this field includes not only development of the new lesion recognition methods [9], but also include a creation of open and publicly available datasets. We are working intensively on extending our own datasets which contain another diseases and findings [7, 8]. The proposed OpenSea system can easily be extend to different diseases by simply using a separate classifier for each

category which will make it easy to run in parallel and more accurate (since it is late fusion and late fusion has been proved as being more accurate [3]). The preliminary results of such a multi-class classification can be found in [10].

It is important to point out that with our method the adjustment of precision and recall is very easy. We can easily increase the recall by using more visual neighbors. This makes it very interesting for the medical use case, because we can get a recall of 1 so that doctors can be sure that we do not miss a true positive example, while still saving them working time because the high precision allows to remove a considerable number of fames.

Possible ways to use the output of the classification tool are presented in the following papers. Here, we use it in a system that allows computer-aided diagnosis. It helps medical experts to find polyps in colonoscopies and also to save medical personnel's working time because they do not have to analyze the whole video. OpenSea has been also used for a system called EIR. This system is built to automatically detect different disease during colonoscopies and capsular endoscopies. The more detailed description of the system can be found in [9, 11, 15]. Different demos of this system have been presented in [12, 16].

Comparing to another existing classification-related software (e.g. Weka[4], a collection of machine learning algorithms for data mining tasks), OpenSea provides not only classification capabilities, but integrates them with feature extraction process. This integration and the simplified data annotation mechanism make the OpenSea tool easy-to-use for all user categories including non-expert users and professionals.

6 CONCLUSION

We presented an easy to use open-source software named OpenSea for image and video classification and showed that the performance regarding processing time and detection accuracy is promising. By making the tool open-source, we hope that we can help other researchers to compare their systems and develop better methods by being able to use it as an easy to get but hard to beat baseline. Moreover, due to the easy way to train the classifier, we hope that also non-experts can use it, especially in the medical use case that we presented. For the future project development, we plan to integrate OpenSea with the latest version of LIRE, speed-up the features extraction process [11], add more features and metrics, integrate custom weights and extend the report generation capabilities.

REFERENCES

[1] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics* 43 (2015), 99–111.

[2] Jorge Bernal, Nima Tajbakhsh, Javier Sanchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjorn Rustad, Ilangko Balasingham, Konstantin Pogorelov, Sungbin Choi, Quentin Debard, Lena Maier-Hein, Stefanie Speidel, Danail Stoyanov, Patrick Brandao, Henry Córdova, Cristina Sánchez-Montes, Suryakanth R. Gurudu, Gloria Fernández-Esparrach, Xavier Dray, Jianming Liang, and Aymeric Histace. 2017. Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge. *IEEE Transactions on Medical Imaging* (2017), 1–19.

[3] Hugo Jair Escalante, Carlos A Hérnadez, Luis Enrique Sucar, and Manuel Montes. 2008. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proc. of ACM ICMR*. 172–179.

[4]https://www.cs.waikato.ac.nz/ml/weka/ [last visited, Feb. 10, 2018]

[4] M. Lux and O. Marques. 2013. *Visual Information Retrieval Using Java and LIRE*. Vol. 25. Morgan & Claypool.

[5] Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405, 2 (1975), 442–451.

[6] Konstantin Pogorelov, Sigrun Losadal Eskeland, Thomas de Lange, Carsten Griwodz, Kristin Ranheim Randel, Håkon Kvale Stensland, Duc-Tien Dang-Nguyen, Concetto Spampinato, Dag Johansen, Michael Riegler, and Pål Halvorsen. 2017. In *Proc. of MMSys*. 112–123.

[7] 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, and Pål Halvorsen. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proc. of MMSYS*. 170–174.

[8] 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, and Pål Halvorsen. 2017. Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proc. of MMSys*. 164–169.

[9] Konstantin Pogorelov, Michael Riegler, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Carsten Griwodz, Peter Thelin Schmidt, and Pål Halvorsen. 2017. Efficient disease detection in gastrointestinal videos - global features versus neural networks. *Multimedia Tools and Applications* 76, 21 (2017), 22493–22525.

[10] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Carsten Griwodz, Thomas de Lange, Kristin Randel, Sigrun Eskeland, Dang Nguyen, Duc Tien, Olga Ostroukhova, and others. 2017. A comparison of deep learning with global features for gastrointestinal disease detection. In *Proc. of CEUR Workshop*.

[11] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Peter Thelin Schmidt, Carsten Griwodz, Dag Johansen, Sigrun L. Eskeland, and Thomas de Lange. 2016. GPU-accelerated Real-time Gastrointestinal Diseases Detection. In *Proc. of CBMS*. 185–190.

[12] Konstantin Pogorelov, Michael Riegler, Jonas Markussen, Mathias Lux, Håkon Kvale Stensland, Thomas Lange, Carsten Griwodz, Pål Halvorsen, Dag Johansen, Peter T Schmidt, and Sigrun L. Eskeland. 2016. Efficient Processing of Videos in a Multi Auditory Environment Using Device Lending of GPUs. In *Proc. of MMSys*. 36.

[13] DMW Powers. 2011. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies* 2, 1 (2011), 37–63.

[14] Michael Riegler, Mathias Lux, Carsten Griwodz, Concetto Spampinato, Thomas de Lange, Sigrun L Eskeland, Konstantin Pogorelov, Wallapak Tavanapong, Peter T Schmidt, Cathal Gurrin, and others. 2016. Multimedia and Medicine: Teammates for Better Disease Detection and Survival. In *Proc. of ACM MM*. 968–977.

[15] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Thomas de Lange, Carsten Griwodz, Peter Thelin Schmidt, Sigrun L. Eskeland, and Dag Johansen. 2016. EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal Endoscopies. In *Proc. of CBMI*. 1–6.

[16] Michael Riegler, Konstantin Pogorelov, Jonas Markussen, Mathias Lux, Håkon Kvale Stensland, Thomas de Lange, Carsten Griwodz, Pål Halvorsen, Dag Johansen, Peter T Schmidt, and Sigrun L. Eskeland. 2016. Computer Aided Disease Detection System for Gastrointestinal Examinations. In *Proc. of MMSys*. 29.

Paper III

Explorative Hyperbolic-Tree-Based Clustering Tool for Unsupervised Knowledge Discovery

Explorative Hyperbolic-Tree-Based Clustering Tool for Unsupervised Knowledge Discovery

Michael Riegler[1], Konstantin Pogorelov[1], Mathias Lux[2], Pål Halvorsen[1], Carsten Griwodz[1]
Thomas de Lange[3], Sigrun Losada Eskeland[4]
[1]Simula Research Laboratory, Norway
[2]Klagenfurt University, Austria
[3]Vestre Viken Hospital Trust, Bærum Hospital, Cancer Registry of Norway, Norway
[4]Department of Medical Research, Bærum Hospital, Vestre Viken Health Trust, Norway

***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.**

I. Introduction

Content-based image retrieval has been an important area of research for quite some time now [1]. A lot of different techniques and methods have been created, and the approaches have become more and more sophisticated. However, there is no one-fits-all approach, and the tools often must be adapted to a particular use-case.

One of the domains we are focusing on is medical images from the human gastrointestinal tract, taken with an endoscope camera inside the body to detect diseases. Even though these images are coming from a particular patient and have been annotated by a particular endoscopist, the domain is not as meta-data rich as intuitively anticipated. Highly trained and specialized medical personnel are scarce human resources, and their priority is on performing medical examinations, not annotating or giving sense to images and videos [2], [3]. Moreover, if videos and frames are shared, the patients personalized information has to be purged from this data or anonymized to ensure privacy of the patients, and especially, in case of shared videos and frames from endoscopic procedures, meta-data is a rare commodity. Therefore, a lot of videos and video frames remain only loosely annotated, and retrieving the images later based on available information is hard.

In this context, we present a prototype mainly designed for visual analysis and annotation of endoscopic images. The prototype application has two main benefits. First, it allows clinical personnel to investigate and analyze vast collections of frames from endoscopic procedures by providing a configurable focus and context view based on frame similarity. Second, it allows for utilizing the focus and context view for annotation and tagging of the dataset, making it more accessible for complementary information systems. While we developed this prototype application for a medical scenario, we strongly believe, and will also show in the evaluation, that it is usable for other scenarios involving interactive browsing, visual analysis or annotation of image or video data. We first investigate the relation between focus and context views and content-based image similarity, as well as discuss the underlying frameworks of the application. We then pick two diverse datasets, one from the medical domain and one from social image collections, to investigate if the proposed abstraction and clustering of the images is applicable through an evaluation. Then, we describe our prototype and show how it can be used to support professional users in the domain of analysis of endoscopic video frames in their daily work routine. Finally, we discuss the contribution of the application and further work on the topic.

II. Related work

Chi [4] defines information visualization in four stages (Table I). First, *raw data* is transformed into an *analytical abstraction*, which is transformed into a *visualization abstraction*, which itself then is presented in a *view*. As indicated in Table I, the data we operate on is images, and for the view stage, we chose a hyperbolic tree visualization.

TABLE I
Prototype stages of visualization and correspondence.

	Stage	**In our prototype**
1	Raw data	Images/ Video frames
2	Analytical abstraction	Image feature descriptors
3	Visualization abstraction	Clusters, centroids and distance values
4	View	Hyperbolic tree

One of the first and most prominent of these approaches was the hyperbolic browser by Lamping, Rao and Pirolli [5]. The underlying idea is, that the visualization abstraction is based on a hierarchy, i.e., a directed tree. In a typical view, the objects would be arranged in a certainly, with those in focus being larger and closer to the center, while those not in

978-1-4673-8695-1/16/$31.00

focus, i.e., the ones being the context, are pushed to the rim of the circle. A hyperbolic view on a hierarchical structure is best described with a fish eye view on a particular tree branch or leaf, with the rest being visible, but out of focus.

The hyperbolic tree visualization is a graph based information visualization strategy [6], which has been applied mostly to data that already closely resembles a tree structure or a directed graph from which a tree can be abstracted including hypertext collections like the WWW, social networks, ontologies and other data where transformation between raw data and abstraction remains on a low complexity level. One of the few examples, where image collections are interpreted as graph structure based on their content, is presented in [7], where the authors employ a force directed placement algorithm to display images on a large video wall. Without the focus and context view, however, the authors are limited by the size of the video wall. Other work of the same authors focuses on displaying images based on content based similarity in a Treemap [8]. The *PhotoTOC* project [9], on the other hand, used clustering to create an *overview+detail view* by clustering images based on color histograms and then presenting the clusters by their medoids. In [10], images are displayed based on their distance with respect to two shape and texture features. Clustering does not take place, but the focus of the visualization lies on the query image and the k nearest neighbors. The rest of the result list is pushed to the outer rim of the visualization providing a context.

III. Analytical and Visual Abstraction

The features for clustering, i.e., the analytical abstraction as defined in Table I, are extracted with LIRE (latest modified version[1]). LIRE supports multiple global and local features out of the box, to allow for easy integration of features in arbitrary applications. Most notable global ones are the Color and Edge Directivity Descriptor (CEDD) [11] as well as the related features including the Joint Composite Descriptor (JCD) [12], the Fuzzy Color and Texture Histogram (FCTH) [13], the Pyramid Histogram of Oriented Gradients (PHOG) [14], the Auto Color Correlogram [15], Local Binary Patterns [16], CENTRIST [17]. Additionally, it includes the MPEG-7 features [18] Edge Histogram, Color Layout and Scalable Color. A detailed description of the extraction process and the features can be found in [19].

For the visualization abstraction stage (see table I), we use WEKA [20]. WEKA is a collection of tools for machine learning and data mining providing also a Java library, which can be directly combined with the LIRE code for our prototype. In the fusion between these two frameworks, LIRE is responsible for the feature extraction and also for the main program logic calling the required functions from WEKA. The coupling allows for optional change of the employed clustering routine. For the experiment described in this paper, the *X-means* clustering algorithm [21] is used, because X-means determines the number of the clusters automatically, which is an important part of the experiment. Our demo also supports K-means and hierarchical clustering [22].

One of the main aspects of our demo is interactivity with the view, i.e., users interact with the created clusters. Clustering, being a well-known technique in machine learning, is used to group entities based on a similarity metric. For instance, images can be group-based on image features (e.g., grouping those with similar colors), or textual user comments can be clustered based on the nouns they contain. For our demo, we use two datasets. One to group pictures showing disease symptoms in a medical scenario, the other to group pictures of the same tagging categories in a social image collection. With visual analysis, these clusters can be investigated by users with domain knowledge about the images content to confirm or reject the grouping within an annotation process.

While being developed for a medical scenario, our prototype is not restricted to a specific domain. Taking advantage of this, we first investigate the appropriateness of the analytical abstraction stage, i.e., the selection of features, as well as the visualization abstraction stage, i.e., the clustering, using two very different publicly available datasets. The first one is the intent dataset of Lux et al. [23]. This dataset contains $1,310$ images crawled from Flickr as well as results from a survey regarding the intentions of the photographers and responses from the photographers as well as crowd-workers judging the images and annotations. The intent categories, from which the users had to choose, are (i) *preserve a good feeling*, (ii) *preserve a bad feeling*, (iii) *show it to family and friends*, (iv) *publish it on-line*, (v) *support a task of mine* and (vi) *recall a specific situation*. For this dataset, the experiment is done for single global features as well as for feature fusions. The second dataset is the ASU-Mayo Clinic polyp dataset which is the biggest publicly available dataset for polyp detection in medical images consisting of 20 videos, with a total number of $18,781$ image frames [24].

On both datasets, we conducted two-step experiments which are slightly different in their final evaluation metric. The first step is clustering the images with our tool based on their global features. The number of clusters is not predetermined, but suggested by X-means. This step is identical for both datasets. For the intent dataset, the mean squared error is then calculated per cluster. In our evaluation, the correlation between the users' feedback and the mean square error of the clusters is computed for the intent dataset. If the correlation coefficient ρ is low, i.e., close to -1, we assume that the method works well, as inter-user-agreement is high while mean square error is low, or the other way around. ρ around 0 or a positive ρ near 1 would indicate that mean square error and user agreement are either not correlated or correlated in the wrong way, implying that the clustering does not work. The intent dataset contains votes of three different users for each category. The users indicates on a 5-point Likert scale how representative an image is for a given category (1, strongly disagree, to 5, strongly agree). For all user votes, the majority vote is calculated and all of them are averaged and normalized.

For the ASU dataset, we can not calculate the mean squared

[1]https://github.com/dermotte/lire, last visited 2016-03-08

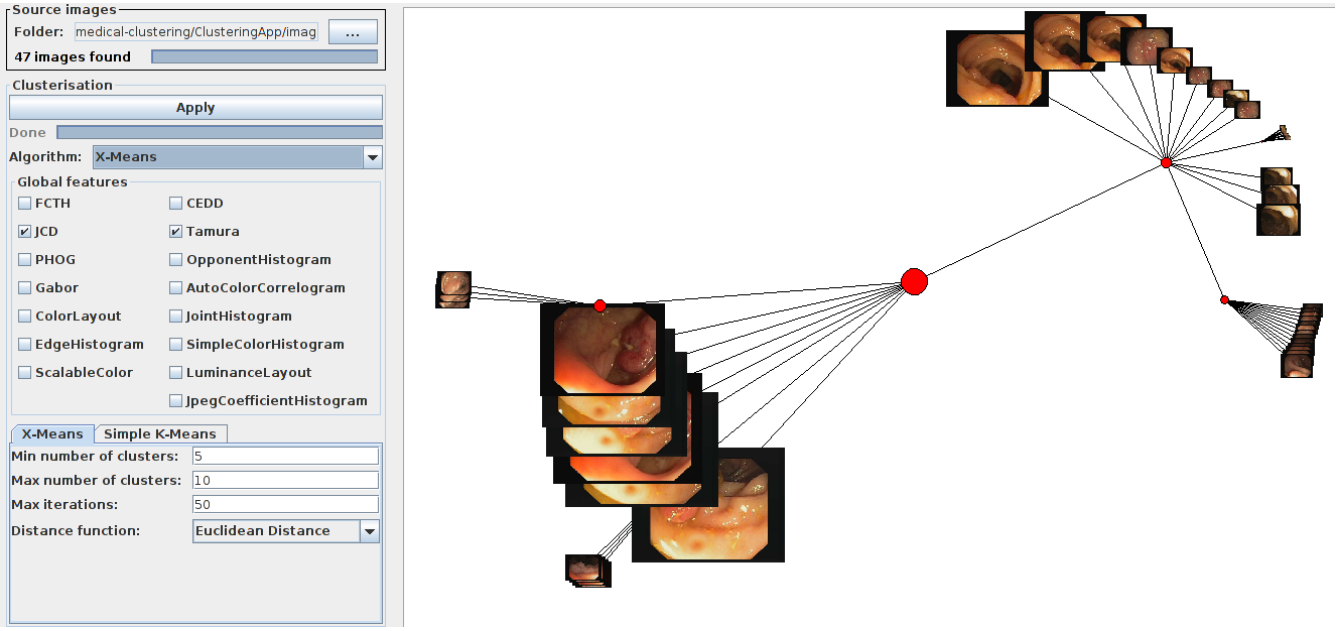


Fig. 1. Demo system: The *left* part contains the settings for the users, and the *right* part shows the output of the clustering as a hyperbolic tree.

error because it contains only binary classification for each frame: a polyp is visible in the image or not. Instead, we calculated the purity of the clusters based on the ground truth provided with the dataset. Furthermore, while we used single global features for the intent dataset, which have been report to work well, we used a combination of the JCD and Tamura features for the ASU dataset. These have been found to work best for this dataset based on an information gain analysis.

Table II shows the results of the experiment based on the intent dataset. As expected, a negative correlation is observed, which means that the clustering results correlate with manual annotations to a degree indicated by the absolute value of ρ. At first, it shows that some global features are more suitable to create clusters that are similar with user judgments than others. For example, FCTH is the best feature for detecting a *publish on-line* intent for an image. A closer look at the clusters generated by FCTH shows that this feature can very well detect if persons are shown in an image, and it seems that most images used for on-line publishing contain one or more persons. Another interesting insight is that semantically similar clusters are also correlated similar to the same feature, e.g., Gabor features for *recall situation* and *preserve good feeling*. This is also an indication that a combination of features is more suitable to provide clusters that are consistent with with user judgments. The last important insight, which is given by this first experiment, is that a simple combination of all features does not automatically lead to better correlation. This indicates that the right choice of feature combinations is important for clustering and that a metric like information gain can give an idea about what features to combine, which we also used in our next experiment. The second experiment with the ASU dataset revealed something similar to the previous experiment. First, we performed information gain analysis to identify the two best features for this dataset. This led us to the features JCD and Tamura, which we combined using early fusion. Based on these features, we performed 4 different tests with different numbers of clusters. We used X-means to determine four clusterings of the dataset. We let X-means determine the number of clusters c for one experiment, then we clustered with $c \in \{2, 4, 100\}$. Based on the created clusters, we calculated the average purity (precision based on the majority class for each cluster). For c equals 2, 4 and 100, we got a purity of 77%, 97% and 95%, respectively. For $c = 234$, the c proposed by the X-means algorithm, the purity is 97%. This indicates that the clustering leads to meaningful results also for the ASU dataset and therefore supports our approach for analytical and visualization abstraction.

TABLE II
CORRELATION ρ BETWEEN MEAN SQUARED ERROR AND USER VOTES FOR DIFFERENT GLOBAL FEATURES OF THE INTENT DATASET [23].

Feature	recall	preserve good	publish	show	support	preserve bad
CEDD	0,165	0,194	0,205	0,285	0,213	-0,05
FCTH	0,085	-0,11	-0,70	-0,32	0,298	-0,27
Gabor	-0,50	-0,40	-0,03	-0,15	-0,08	0,254
Tamura	-0,77	-0,24	0,050	-0,55	0,241	0,517
Luminance Layout	0,060	-0,32	-0,15	-0,30	0,002	0,248
Scaleable Color	0,126	0,295	-0,02	0,060	-0,05	0,094
Opponent Histogram	0,107	-0,07	-0,10	-0,03	0,085	-0,003
AutoColor Correlogram	0,691	0,609	0,739	0,779	-0,47	-0,67
JPEG Coefficent	-0,10	0,006	-0,26	-0,04	-0,48	0,107
Edge Histogram	-0,17	0,643	-0,26	-0,06	-0,51	-0,04
PHOG	-0,52	0,225	0,024	-0,42	0,187	-0,06
JCD	0,168	0,288	0,227	0,193	0,275	-0,26
JointHistogram	0,408	0,262	0,447	0,238	0,396	-0,40
12 Features Combined	-0,14	0,469	-0,11	-0,17	0,215	0,735

IV. PROTOTYPE AND DEMO

Our prototype application combines content-based similarity, unsupervised classification and focus/context views to provide a way to easily explore, analyze and annotate a vast number of video frames or images. Figure 1 shows a screen

shot of the demo application. On the upper left side, users can choose the folder containing the image collection. Below that, the clustering algorithm can be selected. At the moment, we support 3 different algorithms (K-means, X-means and hierarchical clustering). After selecting the clustering algorithm, the application allows to choose one or several different image features. For the screen shot, we limited the list, but the final demo will contain all of the image features provided by LIRE. If more than one feature is picked, they will be combined using early fusion. The final options allow the user to specify the clustering parameters. As a default, we use the values recommended by WEKA. After the users choose the images and all the options, a click on **Apply** creates the clusters and presents them as a hyperbolic tree on the right site. The cluster leaves are represented using the image that is closest to the cluster center, i.e., the cluster medoid. It is possible to interact with the tree by zooming and turning it into different angles. Furthermore, the user can double click on images, which will open the folder containing all images in the selected cluster. A right click on the cluster images allows the user to see information like the cluster center and the purity of the cluster based on the distances. Finally, the users can name/tag the clusters, which adds the tag to the name of the images in the cluster (in this format _"your tag".filetype). For the demo, we will present how our tool works on the two different datasets that we tested here, but we will also have a new large dataset of different endoscopic findings that we will use during the demo presentation.

V. Conclusion

In this paper, we presented a demo application that enables domain experts to use unsupervised clustering algorithms to explore image and video data collections that do not contain meta-data. In the information visualization model of the four stages, the analytical abstraction stage and the visualization abstraction stage correspond to the selection and extraction of image features and the clustering of the feature vectors. We have shown – based on two different datasets – that the clustering leads to good results which correspond to user judgments or ground truth of the datasets, and therefore, provide good candidate methods for the abstraction stages.

For future work, we plan to test the application with domain experts. In our case, endoscopists from two different Norwegian Hospitals. For this test, we already collected a large dataset (200.000 images and 600 videos) from medical procedures. Focus of this user study will be the usefulness of the focus+context view as well as the perceived complexity of the user interface, i.e., the selection of image features and clustering algorithms.

Acknowledgment

This work is funded by the "EONS" FRINATEK project (231687).

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

[2] 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 *Proc. of CBMI*, 2016.

[3] 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 *Proc. of CBMS*. IEEE, 2016.

[4] E. H. Chi, "A taxonomy of visualization techniques using the data state reference model," in *Proc. of IEEE InfoVis*, 2000, pp. 69–75.

[5] J. Lamping, R. Rao, and P. Pirolli, "A focus+ context technique based on hyperbolic geometry for visualizing large hierarchies," in *Proc. of SIGCHI conf. on Human factors in comp. sys.*, 1995, pp. 401–408.

[6] I. Herman, G. Melançon, and M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," *IEEE Trans. on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 24–43, 2000.

[7] Y. Gu, C. Wang, J. Ma, R. J. Nemiroff, and D. L. Kao, "igraph: a graph-based technique for visual analytics of image and text collections," in *IS&T/SPIE Electronic Imaging*, 2015, pp. 939 708–939 708.

[8] C. Wang, J. P. Reese, H. Zhang, J. Tao, and R. J. Nemiroff, "imap: A stable layout for navigating large image collections with embedded search," in *IS&T/SPIE Electronic Imaging*, 2013, pp. 86 540K–86 540K.

[9] J. C. Platt, M. Czerwinski, and B. A. Field, "Phototoc: Automatic clustering for browsing personal photographs," in *Proc. of ICICS-PAM*, 2003, pp. 6–10.

[10] R. S. Torres, C. G. Silva, C. B. Medeiros, and H. V. Rocha, "Visual structures for image browsing," in *Proc. of ACM CIKM*, 2003, pp. 49–55.

[11] S. A. Chatzichristofis and Y. S. Boutalis, "Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval," in *Computer Vision Systems*. Springer, 2008, pp. 312–322.

[12] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux, "Selection of the proper compact composite descriptor for improving content based image retrieval," in *Proc. of IASTED SPPRA*, 2009.

[13] S. Chatzichristofis, Y. S. Boutalis *et al.*, "FCTH: Fuzzy color and texture histogram-a low level feature for accurate image retrieval," in *Proc. of IEEE WIAMIS*, 2008, pp. 191–196.

[14] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. of ACM CIVR*, 2007, pp. 401–408.

[15] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proc. of IEEE CVPR*, 1997, pp. 762–768.

[16] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[17] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[18] S.-F. Chang, T. Sikora, and A. Purl, "Overview of the mpeg-7 standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 688–695, 2001.

[19] M. Lux and G. Macstravic, "The LIRE request handler: A Solr plug-in for large scale content based image retrieval," in *Proc. of MMM*, Dublin, IE, Jan 2014, pp. 374–377.

[20] 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*, vol. 11, no. 1, pp. 10–18, 2009.

[21] D. Pelleg, A. W. Moore *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters." in *ICML*, vol. 1, 2000, pp. 727–734.

[22] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[23] M. Lux, M. Taschwer, and O. Marques, "A closer look at photographers' intentions: a test dataset," in *Proc. of ACM MM workshops - Crowdsourcing for multimedia*, 2012, pp. 17–18.

[24] N. Tajbakhsh, S. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2016.

Paper IV

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

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

Konstantin Pogorelov
Simula Research Laboratory, Norway
University of Oslo, Norway

Michael Riegler
Simula Research Laboratory, Norway
University of Oslo, Norway

Pål Halvorsen
Simula Research Laboratory, Norway
University of Oslo, Norway

Carsten Griwodz
Simula Research Laboratory, Norway
University of Oslo, Norway

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 heterogeneous resources such as GPUs and different optimization techniques.

CCS CONCEPTS

• **Information systems** → **Open source software**; **Multimedia databases**; **Clustering**; Data cleaning; • **Human-centered computing** → **Information visualization**; **Visualization toolkits**; • **Applied computing** → Annotation;

KEYWORDS

Clustering; Visualization; Annotation; Big Collections; Image Browsing; Open Source

ACM Reference format:
Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, and Carsten Griwodz. 2017. ClusterTag: Interactive Visualization, Clustering and Tagging Tool for Big Image Collections. In *Proceedings of ICMR '17, Bucharest, Romania, June 6–9, 2017,* 5 pages.
https://doi.org/http://dx.doi.org/10.1145/3078971.3079018

This work is founded by the Norwegian FRINATEK project "EONS" (#231687).
Contact author's address: Konstantin Pogorelov, Simula Research Laboratory, Oslo, Norway, email: konstantin@simula.no .

ICMR '17, June 6–9, 2017, Bucharest, Romania

ACM ISBN 978-1-4503-4701-3/17/06…$15.00
https://doi.org/http://dx.doi.org/10.1145/3078971.3079018

1 INTRODUCTION

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 softwares, the human eye remains the most important analytic instrument. Research on the topic of interactive image database visualization [8] 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 a lack of tools that implement both fast and efficient image collection visualization together with image content analysis and annotation. In our previous research, we showed [11] the importance of interactive visualization and clustering for unsupervised knowledge discovery in the medical image analysis domain, and we developed the initial application for a medical scenario [6, 12].

To solve the visualization performance issue and provide an efficient solution for visualization and annotation, we continued the development of our tool. We made all necessary modifications and improvements to extend the tool to make it universal and usable for any use case involving interactive browsing, visual analysis and annotation of a large amount of image or video data. In this paper, we present our open source version of the ClusterTag application that we designed for interactive exploration and labeling of big image collections in conjunction with unsupervised image clustering, annotation and tagging.

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

In this paper, we first investigate the state of the art for visualization of big image collection. Next, we describe the clustering methodology, the structure of our system and the optimization approaches implemented in ClusterTag. Then, we describe the open

source software project, its installation requirements, and provide the overall usage instructions. Finally, we discuss the contribution of the application and further work.

2 RELATED WORK

The problem of efficient visual representation of big collections of clustered data is well known. One of the first and most prominent implementations was the hyperbolic browser [3], with the visualization abstraction based on a hierarchy, i.e., a directed tree. A hyperbolic view on a hierarchical structure is best described with a fish eye view on a particular tree branch or leaf, with the rest being visible, but out of focus. A more recent implementation of this approach is presented in [1], where the authors employ a force-directed placement algorithm to display images on a large video wall, showing good performance for a dataset of limited size.

Another promising method for visualization is based on visual similarity of images. In this approach, a combination of automated image content analysis and image cluster visualization is used for building an easy-to-analyze plain representation of an image set. Treemap [14] and Semantic Image Browser (SIB) [15] implement efficient visualization and search in large image databases according to semantic content of images together with the ability to evaluate image annotations through interactive visual exploration of image collections. The disadvantage of Treemap is that it does not support interactive collection browsing. In contrast, SIB does support interactive browsing, but without support for a hierarchical collection structure, which is important for tag-based dataset creation. The approach presented in [7] adds a new visualization scheme that implements a representation of images as various 3D shapes, e.g., a cube or a cone, but without any analyzing, clustering and tagging abilities. The PhotoTOC project [5] uses clustering to create dataset overview and a detailed view representation by clustering images based on color histogram, and then present the clusters using medoids. So far, the PhotoTOC does not provide any collection manipulation functionality. In [13], images are displayed based on their distance with respect to texture features resulting in a 2D dense-visual-representation. Nevertheless, the visualization system does not support both hierarchical structures and multiple clusters within one collection.

Despite the good overall visual representation and easy browsing, methods based on hyperbolic tree visualization are only suitable for small image collections or strictly hierarchical collections with an emphasized tree-structure. For large structured image collections, e.g., in medical image analysis, a different visualization approach is required. Cluster-based image collection representation methods show good browsing capabilities and convenient dataset structure representation together with an additional featured analysis. Nevertheless, the performance of the current implementations of feature extraction, analysis and visualization in the existing frameworks is not good enough to be able to handle large databases containing thousands of images.

3 CLUSTERING

One of the main features of our ClusterTag application is interactivity with a visual collection representation, i.e., users interact with the images and the created or already defined clusters. Clustering, being a well-known technique in machine learning, is used to group entities based on a similarity metric. For instance, textual user comments can be clustered based on the nouns they contain, or images can be grouped based on image features (e.g., grouping those with similar colors).

To help the user in building clusters, we use WEKA[1] [2], a collection of algorithms for machine learning and data mining released as open source software under the GNU General Public License. Clustering algorithms provided by WEKA require the analytical abstraction of image contents. In our approach, we use global features describing image contents in terms of different overall attributes, such as sharpness, color distribution, histogram of brightness, etc. For feature extraction, we use Lucene Image Retrieval (LIRE)[2] [4], which supports extraction and comparison of multiple global and local features out-of-the-box, and allows integration of feature extractors in arbitrary applications. In the fusion between these two frameworks, LIRE is responsible for the feature extraction and processing. WEKA is responsible for unsupervised clustering of images based the extracted features. A detailed description of used global features, corresponding clustering algorithm and clustering performance metrics can be found in [12]. Both the WEKA and LIRE can be easily replaced by other machine learning or feature extraction libraries if desired.

4 VISUALIZATION

To be able to implement a visualization tool for thousands of images simultaneously in real-time and give the user the ability to interact with them, we developed an optimized and highly efficient visualization engine. It is written in Java and uses a number of libraries, enabling high-performance image processing and a real-time image cluster representation.

An overall structure of the software is depicted in figure 1. First, we perform an initial analysis of the image collection file structure. This is the only synchronous operation in the tool, and after it is completed, the user is able to start working with the collection. All other operations on the dataset are asynchronous and implemented as the background threads. The observed initial folder structure is then used in the user interface to draw the visual representations via a painter module. The painter reads the cluster hierarchical structure and interacts with the image in-memory database and the image cache in order to perform an optimized image preloading, rescaling and drawing. This is performed using the hardware accelerated image handler. The database interface performs the in-memory image database synchronization with an on-disk mirror copy. The folder structure update module is 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.

For the image processing tasks, namely loading, saving, rescaling, rotating, masking operations and drawing on top of images, we use the Java bindings of the Open Source Computer Vision Library (OpenCV)[3]. The library provides large number of common and state-of-the-art functions for image processing, with possible acceleration

[1] http://www.cs.waikato.ac.nz/ml/weka/ Accessed: 2017-04-19
[2] http://www.lire-project.net/ Accessed: 2017-04-19
[3] http://opencv.org/ Accessed: 2017-04-19

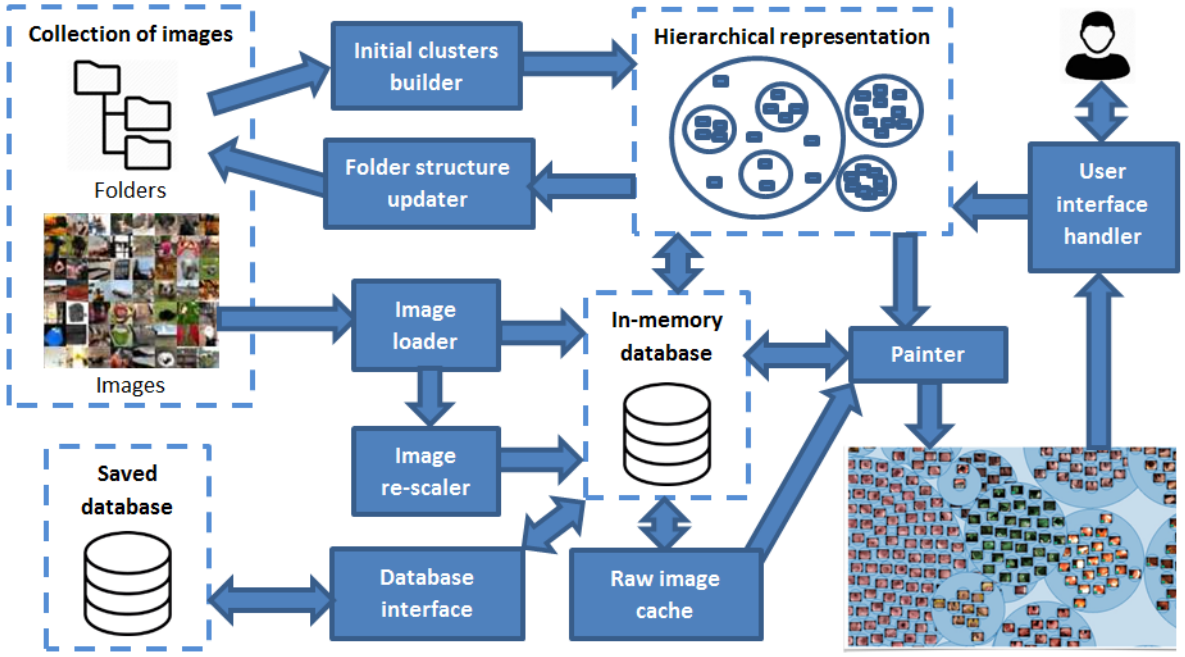


Figure 1: Structure of the visualization and user interface engine of the presented ClusterTag application. A number of caching and intermediate data processing routines are used to make it possible to perform real-time visualization and interaction with huge image collections.

on GPUs. The library is cross-platform and free for use under the open source BSD license.

Real-time drawing of hierarchical representation of the image clusters is implemented using the Lightweight Java Game Library (LWJGL)[4], an open source BSD licensed Java software library for video game developers. It exposes high performance cross-platform libraries, such as OpenGL, Vulkan, OpenAL and OpenCL, and provides a way to get access to high-performance computer resources from Java code.

Regardless of high drawing and image handling performance provided by the OpenCV and LWJGL libraries, several additional optimization techniques have been implemented in the ClusterTag application to allow real-time handling of large image collections. The most important are a database of ready-to-draw pre-processed images, raw image visual representation caching, adaptive image spatial resolution painting, interaction with partially processed collections, multi-threaded image processing and feature extraction (see figure 1 for an overview).

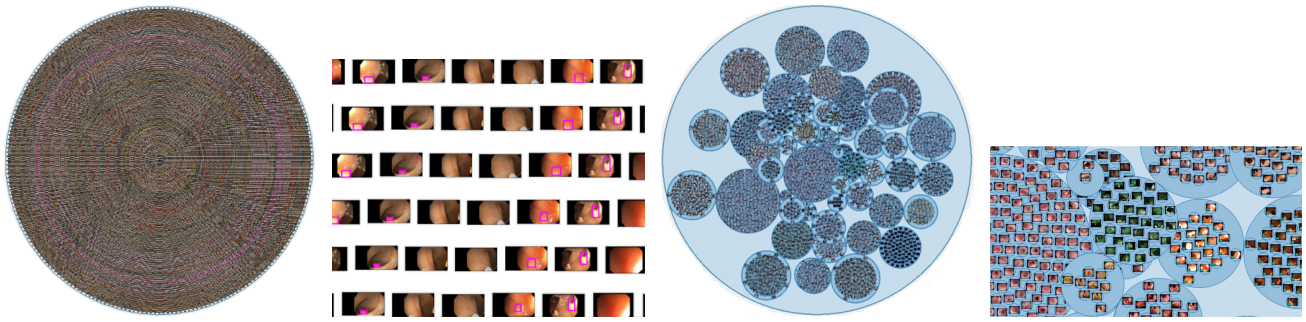
The database of pre-processed images is implemented as a custom binary record storage with a separate index, which allows dynamic database updates with effortless seeking and simultaneous retrieval of multiple records. This is useful for multi-threaded drawing and update of the collection's content at the same time.

The raw image visual representation cache is implemented using concurrent hash maps in conjunction with a custom memory management strategy and is designed to hold the last images used for visual-representation drawing in the memory. Caching is required due to inability to load all the images of the big collection into the system memory. The implemented raw image cache allows fast and smooth image collection browsing without putting the enormous task of image loading and decoding on the CPU.

Our painting engine uses multiple image scales to accelerate the visual-representation drawing. To determine the resolution of the scaled images, we divide the width and height of the original images by power-of-two numbers. The different scaled versions are used depending of the current zoom level of the image collection's field of view. This optimized painting scheme significantly increases the application frame rate, reduces the GPU load and decreases overall system load. All rescaled versions of the image are stored in the database of pre-processed images.

To solve the common slow loading problem of big data collections, we use multiple different techniques in several parts of the system. The main focus is put on the ability to start exploring the image collection as soon as possible 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. Image loading, pre-processing and feature extraction are performed in the background while the user begins to explore the collection and performs fine tuning of the clustering parameters. The visual representation is updated at every screen redraw to reflect the current stage of the collection analysis. All resulting information is immediately stored in the database of pre-processed images and gives the opportunity to continue pre-processing and analyzing the collection at next application startup from the same point.

[4]http://lwjgl.org/ Accessed: 2017-04-19



(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 2: 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.

5 THE CLUSTERTAG PROJECT

The ClusterTag software is an open source project[5] and can be used, modified and distributed under GNU General Public License Version 3. The application has been tested successfully and used with large image collections containing up to 36, 476 images [9, 10].

5.1 Installation

We have tested our software on the Linux, Mac OS X and Windows operating systems. ClusterTag has the following dependencies which have to be downloaded and installed before compilation of the ClusterTag source code: Oracle Java SE Development Kit 8, IntelliJ IDEA 2016.3.4, OpenCV 2.4.13, LIRE 0.9.5 and LWJGL 2.9.3. The detailed compilation and installation instructions can be found on the project's web-page.

5.2 Usage

Our open-source application combines content-based similarity, unsupervised classification and focus/context views to provide a way to easily explore, analyze and annotate a vast number of images or video frames. The application 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 folder structure, and the user can immediately start exploring the collection. Figure 2(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 2(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 2(c) shows a visualization of the collection of medical images clustered using the JCD and Tamura global image features, which produce a number of dense clusters representing visually similar images in the same clusters. The zoomed view of the clustered collection is depicted in figure 2(d).

[5] https://bitbucket.org/mpg_projects/clustertag Accessed: 2017-04-19

The cluster leaves are represented using the image that is closest to the cluster center, i.e., the cluster medoid. It is possible to interact with the view and the clusters by zooming and turning them in different angles. The user can select multiple images to perform grouping. Individual images and image groups can be dragged and dropped between different clusters. The corresponding changes to the file structure of the collection are made in the background. The user can double click on clusters, which opens the folder containing all images in the selected cluster. A right click on the clusters allows the user to see information like the cluster center and the purity of the cluster. Finally, the user can name/tag the clusters, which adds the tag to the name of the images in the cluster. The detailed usage instructions can be found on the project's web-page.

6 CONCLUSION

In this paper, we presented an open source application called ClusterTag which enables users and domain experts to explore, cluster, annotate and tag collections of thousands of images in real-time using an optimized and easy visual representation. We also presented an example of how to use the software with the dataset containing almost 37, 000 medical images.

For future work, we plan to test the application with end users and domain experts. In particular, we will use the application to process an unannotated dataset of endoscopic images from two different Norwegian Hospitals that contains more than 200.000 images and 600 videos. The focus of this user study will be the usefulness of the visual representation, and the perceived complexity of the user interface, i.e., the selection of image features and clustering algorithms as well as performance of our visualization engine. The next development steps for this project will focus on efficient video handling, representation and processing, implementing a plug-in-like feature extraction and clusterisation subsystems, adding new clustering algorithms and unsupervised collection annotation based on preselected image sets.

REFERENCES

[1] Yi Gu, Chaoli Wang, Jun Ma, Robert J Nemiroff, and David L Kao. 2015. iGraph: a graph-based technique for visual analytics of image and text collections. In *Proc. of IS&T/SPIE Electronic Imaging*. 939708–939708.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.
[3] John Lamping, Ramana Rao, and Peter Pirolli. 1995. A focus+ context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proc. of ACM CHI*. 401–408.
[4] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. 2016. LIRE: open source visual information retrieval. In *Proc. of MMSys*. Article no. 30.
[5] John C Platt, Michal Czerwinski, and Brent A Field. 2003. PhotoTOC: Automatic clustering for browsing personal photographs. In *Proc. of ICICS-PAM*. 6–10.
[6] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Peter Thelin Schmidt, Carsten Griwodz, Dag Johansen, Sigrun L. Eskeland, and Thomas de Lange. 2016. GPU-accelerated Real-time Gastrointestinal Diseases Detection. In *Proc. of CBMS*. 185–190.
[7] Marco Porta. 2006. Browsing large collections of images through unconventional visualization techniques. In *Proc. of AVI*. 440–444.
[8] Marco Porta. 2009. New visualization modes for effective image presentation. *International Journal of Image and Graphics* 9, 01 (2009), 27–49.
[9] Michael Riegler, Carsten Griwodz, Concetto Spampinato, Thomas de Lange, Sigrun L. Eskeland, Konstantin Pogorelov, Wallapak Tavanapong, Peter Thelin Schmidt, Cathal Gurrin, Dag Johansen, Håvard Johansen, and Pål Halvorsen. 2016. Multimedia and Medicine: Teammates for Better Disease Detection and Survival. In *Proc. of ACM MM*. 968–977.
[10] Michael Riegler, Konstantin Pogorelov, Sigrun Losada Eskeland, Peter Thelin Schmidt, Zeno Albisser, Dag Johansen, Carsten Griwodz, Pål Halvorsen, and Thomas de Lange. 2017. From Annotation to Computer Aided Diagnosis: Detailed Evaluation of a Medical Multimedia System. *Transactions on Multimedia Computing, Communications and Applications* 9, 4 (2017).
[11] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Thomas de Lange, Carsten Griwodz, Peter Thelin Schmidt, Sigrun Losada Eskeland, and Dag Johansen. 2016. EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal endoscopies. In *Proc. of CBMI*. 1–6.
[12] Michael Riegler, Konstantin Pogorelov, Mathias Lux, Pål Halvorsen, Carsten Griwodz, Thomas de Lange, and Sigrun Losada Eskeland. 2016. Explorative Hyperbolic-Tree-Based Clustering Tool for Unsupervised Knowledge Discovery. In *Proc. of CBMI*. 1–4.
[13] Ricardo S Torres, Celmar G Silva, Claudia B Medeiros, and Heloisa V Rocha. 2003. Visual structures for image browsing. In *Proc. of ACM CIKM*. 49–55.
[14] Chaoli Wang, John P Reese, Huan Zhang, Jun Tao, and Robert J Nemiroff. 2013. iMap: A stable layout for navigating large image collections with embedded search. In *Proc. of IS&T/SPIE Electronic Imaging*. 86540K–86540K.
[15] Jing Yang, Jianping Fan, Daniel Hubball, Yuli Gao, Hangzai Luo, William Ribarsky, and Matthew Ward. 2006. Semantic image browser: Bridging information visualization with automated intelligent image analysis. In *Proc. of IEEE VIS*. 191–198.

Paper V

EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal Endoscopies

EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal Endoscopies

Michael Riegler[*,•], Konstantin Pogorelov[*,•], Pål Halvorsen[*,•], Thomas de Lange[†,♣], Carsten Griwodz[*,•]
Peter Thelin Schmidt[‡,○], Sigrun Losada Eskeland[♣], Dag Johansen[♭]

[*]Simula Research Laboratory, Norway [†]Cancer Registry of Norway [‡]Department of Medicine, Karolinska Institute, Sweden
[•]University of Oslo, Norway [○]Center for Digestive Diseases, Solna and Karolinska University Hospital, Sweden
[♣]Bærum Hospital, Vestre Viken Health Trust, Norway [♭]The Arctic University of Norway, Norway

***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.**

I. Introduction

During the last decades, we have witnessed a paradigm shift where computers and sensors move spatially closer and closer to the user, and we are in the process of moving devices inside the body. In this respect, our scenario is at the intersection of computer science and pathological medicine, where we target a scalable, real-time disease detection system for the gastrointestinal (GI) tract as it is depicted in figure 1. First, we study possible cancer precursors, e.g., polyps, and early cancer detection. Here, we develop both a computer-aided, live analysis system of endoscopy videos and a scalable detection system for screening systems using a wireless video capsule endoscope (VCE), i.e., a small capsule with an image sensor.

In the context of object or pattern detection and tracking in general images and videos, a lot of research has been performed, and current systems are good at detecting human faces, cars, logos, etc. However, detecting diseases in the GI tract is very different from detecting objects like cars. The GI tract can potentially be affected by a wide range of diseases with lesions visible in endoscopy, but findings may also include benign/normal or man-made lesions. The most common diseases are gastric and colorectal cancer (CRC), which are lethal when detected in a late stage (the 5-year survival rate ranges from 93% in stage I to 8% in stage IV [1]). Consequently, early detection is crucial. There are several ways of detecting pathology in the GI tract, but systematic population-wide screening is the most important tool for early detection. However, current methods have limitations regarding sensitivity, specificity, access to qualified medical staff and overall cost.

Fig. 1. The gastrointestinal (GI) tract (Image: kaulitzki/shutterstock.com).

In this scenario, both high precision and recall are of crucial importance, but so is the frequently ignored system performance that can provide feedback in real time. The most recent and most complete related work is the polyp detection system Polyp-Alert [2], which can provide near real-time feedback during colonoscopies. However, it is limited to polyp detection, and it is not fast enough for live examinations. To further aid and scale such examinations, we present EIR[1], an efficient and scalable automatic analysis and feedback system for medical data like videos and images. The system supports endoscopists in the detection and interpretation of diseases in the GI tract. EIR has initially been tested in scenarios supporting endoscopists in detection and interpretation of potential diseases in lower portions of the GI tract (large bowel). However, the main objective is to automatically detect abnormalities in the whole GI tract. Therefore, the aim is to develop both (i) a live system assisting the visual detection of, for example, polyps during colonoscopies and (ii) a future fully automated screening of the GI tract using VCEs. Both aims impose strict requirements on the accuracy of the detection to avoid false negative examinations (overlooking a disease) as well as low resource consumption. The live-assisted system also introduces a real-time processing requirement (defined as being able to process at least 30 frames or images per second). In this paper, the initial framework of our complete system is presented. To detect mucosal lesions in the colon, we built a system

[1]In Scandinavian mythology, EIR is a goddess with medical skill.

978-1-4673-8695-1/16/$31.00

combining filters using machine learning, image recognition and extraction and comparison of global and local image features. Furthermore, it is easy to add new filters or other types of data, such as patient records or sensor data, to increase accuracy or enable detection of other pathologies. Moreover, we evaluate our prototype by training classifiers that are based on different image recognition approaches. It is important to point out that these classifiers can also process other input like sensor data. We also test the generated classifiers with different data and thereby evaluate the different approaches for feasibility of colonic polyp recognition and localisation. The initial results from our experimental evaluation show that, (i) the detection and localisation accuracy can reach the same performance or outperform other current state-of-the-art methods and (ii) the system performance can reach real-time in terms of video processing up to high definition resolutions. Additionally, it is extensible with more data and diseases thorough parallel detection at run time. The rest of the paper is organized as follows: Firstly, in section II, we briefly introduce our medical case study. Next, we present related work in the field and compare it to the presented system in section III. This is followed by presenting the complete system in section IV. After that we, present an evaluation of the system in section V, and in section VI we discuss two cases where our system will be used in two medical examinations by our collaborators. Finally, we conclude with section VII.

II. Gastrointestinal Endoscopy

The GI tract illustrated in figure 1 can potentially be affected by various abnormalities and diseases, e.g., CRC, a major health issue world wide. Early detection of CRC or polyps as predecessors of CRC is crucial for survival, and several studies demonstrate that a population-wide screening program improves the prognosis and can even reduce the incidences of CRC [3]. As a consequence, in current European Union guidelines, screening for colorectal cancer is recommended for the age group over 50 [4]. 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 positive screening tests [5]. However, endoscopies are invasive procedures and may lead to great discomfort for patients. Extensive training of physicians or nurses is required to perform the examination. They are performed in real-time and therefore challenging to scale to a large population. Additionally, the procedure is expensive. In the US, for example, colonoscopy is the most expensive cancer screening process, with annual costs of 10 billion dollars ($1,100$\$-$6,000$\$/person) [6], and with a time consumption of about one medical-doctor-hour and two nurse-hours per examination. As a first step, we target the detection of colorectal polyps, which are known precursors of CRC (see for example figure 2). 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 prevents the development of cancer and the risk of getting CRC in the following 60 months after a colonoscopy depend largely on the endoscopist's ability to detect polyps [7]. Nevertheless, our system will be extended to support detection of multiple abnormalities and diseases of the GI tract by training the classifiers using different datasets.

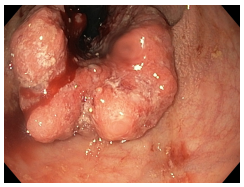


Fig. 2. Colorectal cancer that can be found using colonoscopy.

III. Related Work

Detection of diseases in the GI tract has mostly focused on polyps. This is most probably due to the lack of data in the medical field and polyps being a condition with at least some data available. However, none of the related work is able to do real-time detection or support doctors by computer-aided diagnosis during colonoscopies in real-time. Furthermore, all of them are limited to a very specific use case, which in the most cases is polyp detection for a specific type of camera. Table I gives an overview of the best working methods.

As one can see in Table I, several algorithms, methods and partial systems have been proposed and have, at first glance, achieved promising results in their respective testing environment. However, 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 dataset and test scenarios. Therefore, overfitting for the specific datasets can be a problem and can lead to unreliable results.

The first approach from Wang et al. [2] is the most recent and best-working one in the field of polyp detection. A list of more related work can be found in their paper. Polyp-Alert [2] is able to give near real-time feedback during colonoscopies. The system can process 10 frames per second and uses visual features and a rule-based classifier to detect the edges of polyps. Further, Polyp-Alert distinguishes between clear frames and polyp frames in its detection. The researchers report a performance of 97.7% correctly detected polyps, based on their dataset, which consists of 52 videos taken from different colonoscopes. Unfortunately, the dataset is not publicly available, and therefore, a detection performance comparison is not possible. Since neural networks (NN) are commonly used nowadays, they are also discussed in relation to the GI tract analysis. We identified two main points that make NNs less useful for our use case [17]. Firstly, (i) their training requires a lot of good training data, which is a big a problem in the medical field [18], and (ii) NNs are not easy to design for probabilistic results, which is important to support medical doctors during decision making [19].

In summary, a lot of good related work with interesting approaches for polyp detection exists. However, 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

TABLE I
A PERFORMANCE COMPARISON OF POLYP DETECTION APPROACHES. NOT ALL PERFORMANCE MEASUREMENTS ARE AVAILABLE FOR ALL METHODS, BUT INCLUDING ALL AVAILABLE INFORMATION GIVES AN IDEA ABOUT EACH METHOD'S PERFORMANCE.

Publ./System	Detection Type	Recall / Sensitivity	Precision	Specificity	Accuracy	FPS	Dataset Size
Wang et al. [2]	polyp / edge, texture	97.70%	–	–	95.70%	10	1.8m frames
Wang et al. [8]	polyp / shape, color, texture	81.4%	–	–	–	0.14	1,513 images
Mamonov et al. [9]	polyp / shape	47%	–	90%	–	–	18,738 frames
Hwang et al. [10]	polyp / shape	96%	83%	–	–	15	8,621 frames
Li and Meng [11]	tumor / textural pattern	88.6%	–	96.2%	92.4%	–	–
Zhou et al. [12]	polyp / intensity	75%	–	95.92%	90.77%	–	–
Alexandre et al. [13]	polyp / color pattern	93.69%	–	76.89%	–	–	35 images
Kang et al. [14]	polyp / shape, color	–	–	–	–	1	–
Cheng et al. [15]	polyp / texture, color	86.2%	–	–	–	0.076	74 images
Ameling et al. [16]	polyp / texture	AUC=95%	–	–	–	–	1,736 images
EIR-system	abnormalities/30 features	98.50%	93.88%	72.49%	87.70%	30-65	18,781 frames

scenario and; (iii) provide a 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.

IV. EIR BASIC IDEA

Our objective is to develop a system that supports doctors in disease detection in the GI tract. The system must (i) be easy to use and less invasive for the patient that existing methods, (ii) be easy to extend to different diseases, (iii) handle of multimedia content in real time, (iv) be usable for real-time computer-aided diagnosis, (v) achieve high classification performance with minimal false-negative classification results and (vi) have a low resource consumption. These properties potentially provide a scalable system with regard to cost, medical specialists required for a larger population, and number of users potentially willing to be screened. Therefore, EIR consists of three parts: The annotation subsystem, the detection and automatic analysis subsystem and the visualization and computer-aided diagnosis subsystem.

A. Annotation Subsystem

The purpose of the annotation subsystem is the efficient collection of training data for the detection and automatic analysis subsystem. It is well known that training data is very important for a good classification system. Nevertheless, in the medical field, the time of the experts and access to multimedia data are two resources that are quite limited. This is primarily because of high everyday workload for physicians, but also due to legal issues. For each image or video, patient consent has to be collected before research can be done, making it a very cumbersome task. Moreover, the annotation of videos itself is very time-consuming, and the quality of annotations depends on the experience and concentration of the physicians [20]. For example, in a VCE procedure, there are about $216,000$ images per examination, and a very experienced endoscopist needs at least 60 minutes to view and analyse all the video data [21]. Due to this limitation, it is important to develop automatic methods that can reduce the burden on physicians and speed up the screening process. We therefore developed an efficient semi-automatic annotation subsystem [22]. This annotation system is the entry point into our whole system. Since the medical doctor is usually located in a hospital with restrictions to data security, the implementation of the software is done with standard web technologies, which do not require any installation on the hospital's systems. This includes the storing of all information on the system-side and moves the responsibility of maintaining the system and the data integrity from the user to the system. Besides getting data for the EIR system to enable automatic screening, the annotation subsystem makes it possible to use the annotated videos in a medical video archive for documentation or teaching purposes.

B. Detection and Automatic Analysis Subsystem

These subsystems for algorithmic analysis are designed in a modular way, so that they can be extended to different diseases or subcategories of disease, as well as other tasks like size determination, etc. At the moment, this subsystem consists of two parts, the detection subsystem that detects irregularities in video frames and images and the localisation subsystem that localizes the exact position of the disease. The detection can not determine the location of the found irregularity. The location determination is done by the localisation subsystem. The localisation subsystem uses the output of the detection system as input.

1) Detection Subsystem: This part of the system is not designed to detect the precise position of an abnormality like a polyp or bleeding, but rather to detect whether there is something in the current frame of the video or not. All the frames that we process can be separated into two disjoint sets which can also be seen as the model for the detector. These two sets contain example images for abnormalities and images without any abnormality. Each of these sets can be seen as the model for a specific disease. The detection system is built in a modular way and can easily be extended with new models. This would for example allow to first detect a polyp and then to distinguish between a polyp with low or high risk to developing into CRC by using the *NICE* classification[2]. To compare and determine the abnormalities in a given video frame, we use global image features, i.e., because they are easy and fast to calculate, and because the exact position is at this point of the system not needed. In previous work, we showed that global features can indeed outperform or at least reach the same results as local features [23]. The basic idea is based

[2]http://www.wipo.int/classifications/nice/en/

on an improved version of a search based method for image classification presented in [23]. We create the indexes from visual features extracted from video frames or images. However, the number of needed examples is rather low compared to other methods. The index also contains information about the presence and type of any disease in the frame or image. A classifier can then search the index for the frames that are most similar to a given input frame. Based on the classification of the results, the detection subsystem then decides which abnormality the input frame belongs to. The whole detector is realised with two separate tools, an indexer and a classifier. We have released the indexer and the classifier as a separate project called *OpenSea*[3]. The computational nature of the indexing part is similar to what we know as batch processing. Therefore, creating the models for the classifier could be done off-line, and it is not influencing the real-time capability of the system, because it is only done once at the very first time when the training data is inserted into the system. Visual features to calculate and store in the indexes can be chosen based on the abnormality because, for different types of disease different set of features or combinations are better. 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 subsystems model consists of. The classifier uses indexes generated by the indexer described before. 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 allows to choose how many CPU cores are used. Ongoing work includes to port parts of the system to GPUs to further increase the performance.

2) Localisation Subsystem: The localisation subsystem is intended for exact positioning of a lesion, which is used to show markers on the frame or image containing the disease. This information is then used in the visualization subsystem. All images that we process during the localisation step come from the positive frames list generated by the detection subsystem. Processing of the images is implemented as a sequence of intraframe 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 itself 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 and ambient light. Moreover, the image itself can be interleaved, noisy, blurry and over or under exposed, and it can contain borders and subimages. Apart from that, it can have various resolutions depending on the type of endoscopy equipment used. Endoscopic images usually have a lot of flares and flashes caused by high power light source located close to the camera. All these nuances affect the local features detection methods negatively and have to be specially treated to reduce localisation precision impact. In our case, several, sequentially applied filters are used to prepare raw input images for the following analysis. These analyses are RGB to YCbCr color space conversion, borders and subimages removing, flares masking and low-pass filtering. After the pre-filtering, the images are used for further analysis.

At the moment, we have implemented the detection of colon polyps using our local features approach. The main idea of this localisation algorithm is to use the polyps' physical shape to find the exact position in the frame. In most cases, the polyps have the shape of a hill located on 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 those types of objects, we use the following sequence of filters: binary noise reduction filter, 2D-gradient filter, threshold borders detection filter and binary noise removing filter. The next step creates a filtered binary image approximated by a set of ellipses from which we build energy maps based on the ellipse's size and border points precision approximation and matching. The final coordinates of one or more polyps in the frame are chosen by looking for the maximum in the energy map.

C. Visualization and Computer Aided Diagnosis Subsystem

This subsystem has two main purposes. 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. Therefore, we have the TagAndTrack subsystem [22] that can be used for visualization and computer-aided diagnosis. We developed a web technology-based visualization that can be used to support medical experts while being very easy to use and distribute. This tool simply takes the output of the systems detection and localisation part and creates a web-based visualization, which can then be combined with a video sharing platform where doctors are able to watch, archive, annotate and share information.

V. Evaluation

For all of the subsequent measurements, we used the same computer (32 AMD CPU cores Linux server, 128GB ram). It is important to point out that the used hardware is quite old (ca. 4 years). Newer hardware would most probably lead to better performance for all the tests, but the evaluation shows that even on old hardware the system performs as intended. For all experiments, we used the ASU-Mayo Clinic polyp database[4]. This is currently the biggest publicly available dataset consisting of 20 videos from standard colonoscopies (converted from WMV to MPEG-4 for the experiments) with a total of 18,781 frames and different resolution up to full

[3]https://bitbucket.org/mpg_projects/opensea

[4]http://polyp.grand-challenge.org/

HD [24]. For the detection and localisation accuracy, we used the common standard metrics precision, recall/sensitivity and F1 score. We conducted a leave-one-out cross-validation to evaluate this part of the system, which is a method that assesses the generalization of a predictive model. In our case, it describes the process where the training and testing datasets are rotated, leaving out a single different non-overlapping item or portion for testing, and using the remaining items for training. This process is repeated until every item or portion has been used for testing exactly once [25].

EIR allows us to use several different global image features for the classification. The more image features we use, the more computationally expensive the classification becomes. Further, not all image features are equally important or provide equally good results for our purpose. As a first step, we therefore need to find out which image features we want to use for classification. In order to understand which image features provide the best results, we generated indexes containing all possible image features for all frames of all video sequences from the ASU-Mayo Clinic database. These indexes can be used for several different measurements and also for leave-one-out cross-validation. Using our detection system, the built-in metrics functionality can provide information on the performance of different image features for benchmarking. Further, it provides us with separate information for every single image feature, as well as the late fusion of all the selected image features. All used features are described in detail in [26].

Accuracy. Based on the evaluation of different combinations of image features using 30 different features and information gain analysis, the image features JCD and Tamura were identified to be the best ones for polyp detection. The last row of table I shows our approaches' performance to give a comparison. We achieve an average precision of 0.9388, an average recall of 0.9850, and an average F1 score value of 0.9613. In other words, the results mean that we can detect polyps with a precision of almost 94%, and we detect almost 99% of all polyp containing frames. If we compare this to the best performing system in table I, it seems that Polyp-Alert reaches slightly higher detection accuracy. But, our system is faster and can detect polyps in real-time. Furthermore, our system is not designed and restricted to detect only polyps, and can be expanded to any possible disease if we have the correct training data.To evaluate the performance of the localisation subsystem we used the exact positions of the polyps as provided in the ASU-Mayo clinic polyp database as ground truth. Overall, we reached for the localisation an average precision of 0.3207, a recall of 0.3183 and a F1 score of 0.3195.

Speed. We also performed some initial system performance tests. For all these tests, we used 3 videos from 3 different endoscopic devices and different resolutions. The three videos have the resolutions $1,920x1,080$, $856x480$ and $712x480$. We chose these videos to show the performance under different requirements that the system will have to face when it is used. As figure 3 shows, EIR reaches the required 30 frames per second with 16-26 CPUs. This is true for all three videos that we used. For the future an implementation using GPUs will be important to cope with the high number of needed cores.

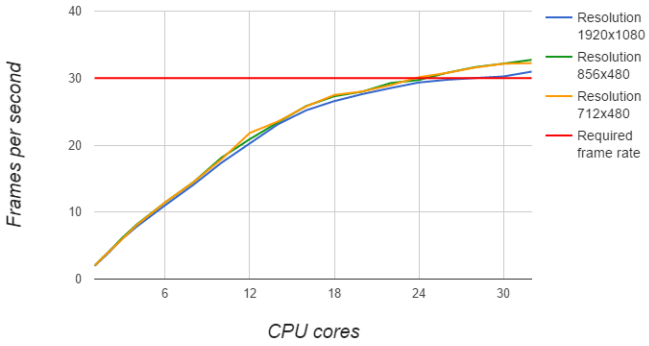


Fig. 3. Processing performance in frames per second.

VI. Real World Use Cases

In this section, we will describe two real world use cases where the presented system can be used. The first one is a live system that will support medical doctors during endoscopies. Currently, we are working on setting it up in one of our partner hospitals. The second one is a system that will automatically analyse videos captured by VCEs. Several hospitals all over Europe and US are involved in this part, and currently, we are collecting data. 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.

Live System. Endoscopy is a common gastrointestinal examination and is essential for the diagnosis of most mucosal diseases in the gastrointestinal tract, particularly diagnosis of CRC and its precursors. Previous studies have demonstrated that a major challenge is the detection rate of lesions [27]. 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 analyses the video frames that are recorded by the colonoscope. At the beginning, we plan to optically show the physician (for example with a red or green frame around the video) when the system detects something abnormal in the actual frame or not. This can also be extended to the determination of what disease the system most probably detected and provide this information to the doctor. Apart from supporting the medical expert during the colonoscopy, the system can also be used to document the procedure. After the colonoscopy, an overview can be given to the doctors where they can make changes or corrections, and add additional information. This can then be stored for later purposes or used as a written endoscopy report. A demo of the live system is presented and described in [28]

Wireless Video Capsule Endoscope. The present VCEs have a resolution of around $256x256$ with 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 difficult use the images. Nevertheless ongoing work tries to improve the state-of-the-art technology which will make it possible to use the methods and algorithms developed for colonoscopies also for VCEs [29]. The multi-sensor VCE

is swallowed in order to visualize the GI tract for subsequent diagnosis and detection of GI diseases. Thus, people may be able to buy VCEs at the pharmacy, and connect and deliver the video stream from the GI tract to the phone over a wireless network. The video footage can be processed in the phone or delivered to our system, which finally analyses the video automatically. 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.

VII. Conclusion

In this paper, a multimedia system for disease detection and classification in the GI tract has been presented. We briefly described the whole pipeline of the system from annotation (data collection for system learning) to visualization (doctor feedback). A detailed evaluation in terms of detection and localisation accuracy and system performance has been performed. These experiments showed that the proposed system can achieve equal results to state-of-the-art methods in terms of detection accuracy for state-of-the-art endoscopic data. Further, we showed that the 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 also presented automatic analysis of VCE videos and live support of colonoscopies as two real-world use cases that will benefit from the proposed system and will actually be tested and used in our partner hospitals. For future work, we plan to improve the detection and localisation accuracy of the system and include more different abnormalities to detect. Presently, we are working with medical experts to collect more training data. Additionally, we currently work on the set-up of the real-world use cases in the hospitals. Finally, to further improve the performance of the system, we work on an extension that allows the system to use GPUs to further utilize the parallelization potential of the workload [30].

Acknowledgment

This work is founded by the FRINATEK project "EONS" #231687.

References

[1] J. B. O'Connell, M. A. Maggard, and C. Y. Ko, "Colon cancer survival rates with the new american joint committee on cancer sixth edition staging," *NCI*, vol. 96, 2004.

[2] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. de Groen, "Polyp-alert: Near real-time feedback during colonoscopy," *CMBP*, no. 3, 2015.

[3] O. Holme, M. Bretthauer, A. Fretheim, J. Odgaard-Jensen, and G. Hoff, "Flexible sigmoidoscopy versus faecal occult blood testing for colorectal cancer screening in asymptomatic individuals," *CDSR*, 2013.

[4] L. von Karsa, J. Patnick, and N. Segnan, "European guidelines for quality assurance in colorectal cancer screening and diagnosis. first edition–executive summary," *Endoscopy*, 2012.

[5] S. Mallery and J. Van Dam, "Advances in diagnostic and therapeutic endoscopy," *Med Clin North Am*, vol. 84, no. 5, pp. 1059–83, 2000.

[6] The New York Times, "The $2.7 Trillion Medical Bill," http://goo.gl/CuFyFJ, [last visited, Nov. 29, 2015].

[7] 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," *JM*, vol. 362, no. 19, pp. 1795–1803, 2010.

[8] 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," *In proc. of BHI*, vol. 18, no. 4, pp. 1379–1389, 2014.

[9] A. Mamonov, I. Figueiredo, P. Figueiredo, and Y.-H. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1488–1502, July 2014.

[10] 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*, Sept 2007, pp. 465–468.

[11] 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*, vol. 16, no. 3, pp. 323–329, May 2012.

[12] 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*, Oct 2014, pp. 237–241.

[13] L. A. Alexandre, J. Casteleiro, and N. Nobreinst, "Polyp detection in endoscopic video using svms," in *Proc. of PKDD*, 2007, pp. 358–365.

[14] J. Kang and R. Doraiswami, "Real-time image processing system for endoscopic applications," in *Proc. of CCECE*, vol. 3, 2003, pp. 1469–1472.

[15] 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 *In proc. of MDAISM*. Springer, 2008, pp. 62–72.

[16] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, "Texture-based polyp detection in colonoscopy," in *Proc. of BfM*. Springer, 2009, pp. 346–350.

[17] C. Chin and D. E. Brown, "Learning in science: A comparison of deep and surface approaches," *Journal of Research in Science Teaching*, vol. 37, no. 2, pp. 109–138, 2000.

[18] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[19] D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, no. 1, pp. 109–118, 1990.

[20] B. Giritharan, X. Yuan, J. Liu, B. Buckles, J. Oh, and S. J. Tang, "Bleeding detection from capsule endoscopy videos," in *Proc. of EMBS*, 2008.

[21] B. Li and M. Q. H. Meng, "Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments," *CBM*, vol. 39, no. 2, pp. 141–147, 2009.

[22] 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*, ser. MMSys '15. New York, NY, USA: ACM, 2015, pp. 73–76. [Online]. Available: http://doi.acm.org/10.1145/2713168.2713184

[23] 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 *In proc. of MM*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 397–406. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654894

[24] N. Tajbakhsh, S. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," 2015.

[25] B. Efron and R. Tibshirani, "Improvements on cross-validation: The .632+ bootstrap method," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. pp. 548–560, 1997. [Online]. Available: http://www.jstor.org/stable/2965703

[26] M. Lux and O. Marques, *Visual Information Retrieval Using Java and LIRE*. Morgan & Claypool, 2013.

[27] T. de Lange, S. Larsen, and L. Aabakken, "Image documentation of endoscopic findings in ulcerative colitis: photographs or video clips?" *GE*, vol. 61, no. 6, pp. 715–720, 2005.

[28] M. Riegler, K. Pogorelov, J. Markussen, M. Lux, H. K. Stensland, T. de Lange, C. Griwodz, P. Halvorsen, D. Johansen, P. T. Schmidt, and S. L. Eskeland, "Computer aided disease detection system for gastrointestinal examinations," in *Proc. of MMSys*, 2016.

[29] A. Khaleghi and I. Balasingham, "Wireless communication link for capsule endoscope at 600 mhz," in *Proc. of EMBC*, 2015, pp. 4081–4084.

[30] 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 *Proc. of CBMS*, 2016.

Paper VI

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

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

MICHAEL RIEGLER and KONSTANTIN POGORELOV, Simula Research Laboratory and University of Oslo
SIGRUN LOSADA ESKELAND, Bærum Hospital, Vestre Viken Hospital Trust
PETER THELIN SCHMIDT, Karolinska Institutet, Department of Medicine, Solna and Karolinska University Hospital, Center for Digestive Diseases, Stockholm
ZENO ALBISSER, Simula Research Laboratory and University of Oslo
DAG JOHANSEN, UiT - The Arctic University of Norway
CARSTEN GRIWODZ and PÅL HALVORSEN, Simula Research Laboratory and University of Oslo
THOMAS DE LANGE, Bærum Hospital, Vestre Viken Hospital Trust and Cancer Registry of Norway

Holistic medical multimedia systems covering end-to-end functionality from data collection to aided diagnosis are highly needed, but rare. In many hospitals, the potential value of multimedia data collected through routine examinations is not recognized. Moreover, the availability of the data is limited, as the health care personnel may not have direct access to stored data. However, medical specialists interact with 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 article, we present a novel, holistic multimedia system aiming to tackle automatic analysis of video from gastrointestinal (GI) endoscopy. The proposed system comprises the whole pipeline, including data collection, processing, analysis, and visualization. It combines filters using machine learning, image recognition, and extraction of global and local image features. The novelty is primarily in this holistic approach and its real-time performance, where we automate a complete algorithmic GI screening process. We built the system in a modular way to make it easily extendable to analyze various abnormalities, and we made it efficient in order to run in real time. The conducted experimental evaluation proves that the detection and localization accuracy are comparable or even better than existing systems, but it is by far leading in terms of real-time performance and efficient resource consumption.

CCS Concepts: • **Information systems → Multimedia information systems;**

Additional Key Words and Phrases: Medical multimedia system, gastrointestinal tract, evaluation

ACM Reference Format:
Michael Riegler, Konstantin Pogorelov, Sigrun Losada Eskeland, Peter Thelin Schmidt, Zeno Albisser, Dag Johansen, Carsten Griwodz, Pål Halvorsen, and Thomas de Lange. 2017. From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system. ACM Trans. Multimedia Comput. Commun. Appl. 13, 3, Article 26 (May 2017), 26 pages.
DOI: http://dx.doi.org/10.1145/3079765

This work is founded by the Norwegian FRINATEK project "EONS" (#231687).
Authors' addresses: M. Riegler, K. Pogorelov, Z. Albisser, C. Griwodz, and P. Halvorsen, Simula Research Laboratory, P.O.Box 134, 1325 Lysaker, Norway; emails: {michael, konstantin, zeno, griff, paalh}@simula.no; S. L. Eskeland, Bærum Hospital, Sogneprest Munthe-Kaas vei 100,1346 Gjettum; email: sigesk@vestreviken.no; P. T. Smidt, Karolinska University Hospital, Solna 171 76 Stockholm, Sweden; email: peter.thelin-schmidt@karolinska.se; D. Johansen, University of Tromsø, Postboks 6050 Langnes, 9037 Tromsø; email: dag@cs.uit.no; T. de Lange, Cancer Registry of Norway, Postboks 5313 Majorstuen, 0304 Oslo; email: Thomas.de.Lange@kreftregisteret.no.

ACM 1551-6857/2017/05-ART26 $15.00
DOI: http://dx.doi.org/10.1145/3079765

1. INTRODUCTION

Devices such as sensors and cameras have become much smaller in the last years. Literally, some of the devices, like cameras, have been moved inside the human body. Thus, there has for some time been a move toward an interdisciplinary research area that combines the medical and multimedia research fields [10, 22, 58]. In particular, for reasons like disease severity, cost, personnel time consumption, and examination scalability, there is a need to develop a real-time and scalable abnormality detection system for videos from gastrointestinal (GI) endoscopy examinations. In this respect, one should target an analysis system for endoscopies that can be used for both a live computer-aided diagnosis system and a scalable detection system for a novel in-line screening system using wireless video capsule endoscopes (VCEs).

The GI tract can potentially be affected by a wide range of diseases. For example, three of the six most common cancer types are located in the GI tract, with about 2.8 million new luminal GI cancers (esophagus, stomach, colorectal) yearly and a mortality of about 65% [64]. These diseases, as well as benign findings or man-made (iatrogenic) lesions are frequently visualized with endoscopes. Gastric- and colorectal cancer are the most common cancers and lethal when detected in late stages. Consequently, early detection is crucial. There are several ways of detecting pathology in the GI tract, and regular systematic screening of the population cohort (everyone above 50 years) is the most important tool for early detection and even cancer prevention. However, current methods have limitations regarding sensitivity, specificity, access to qualified medical staff and overall cost.

To aid and scale endoscopic examinations, we have developed EIR, named after a Goddess with medical skills in Scandinavian mythology. EIR is an end-to-end efficient and scalable information retrieval system for medical data like videos and images, sensor data, and patient records, i.e., EIR combines a content-based similarity search with statistical classifiers from the training data. The system supports endoscopists in the detection and interpretation of diseases in the GI tract but can basically be expanded to any other use-case. The main objective is to automatically detect abnormalities in the whole GI tract. Therefore, the aim is to develop both (i) a live system assisting the visual detection of, for example, polyps during colonoscopies and (ii) a future fully automated first line screening for GI diseases using VCEs. Both aims pose strict requirements for the accuracy of the detection in order to avoid false-negative findings (missing a disease) as well as low resource consumption. The live assisted system also introduces a real-time processing requirement. In this article, following some of ACM multimedia (MM) brave new ideas [44], we extend our initial work on EIR [45] to include a more detailed description of our improved sub-systems. Therefore, the main contributions are presenting the copious improvements of the different sub-systems, an in-depth evaluation of global features' detection accuracy, and a new extensive performance evaluation analyzing system execution time and memory consumption. Furthermore, we provide an evaluation of the effect of the amount of available training data and an accuracy performance comparison with other systems - both at a grand challenge for endoscopic video analysis and against systems found in literature. An important design decision has been to build on state-of-the-art sub-component solutions in our quest to find an optimal complete end-to-end system meeting both accuracy and performance requirements. Thus, our focus has not been on improving sub-components in isolation, but rather providing an integrated system that more or less can be put to good use in the next phase.

Although our system is not limited to one single disease, detecting abnormalities and diseases in the GI tract is very different from detecting objects like, for example, cars, people, or buildings, which have been the focus for most existing research. Our initial experiments target a scenario where we detect colorectal polyps, a potential precursor for colorectal cancer (CRC). Statistics show that the lifetime risk of getting CRC, the

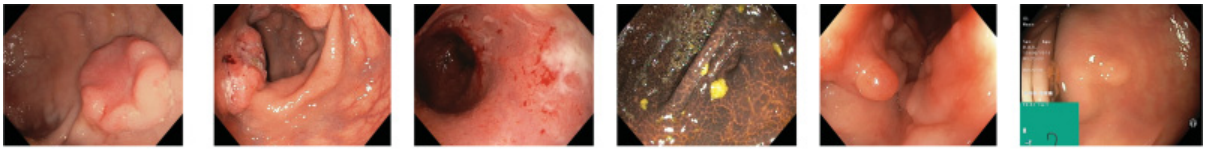


Fig. 1. An inconclusive list of abnormalities that can be found using colonoscopy.

second most common cancer for both genders, is 6% [15], and a previous trial has shown that CRC may be prevented by polyp removal [47]. Obviously, both high precision and recall are of crucial importance, but so is the often ignored system performance in order to provide live feedback and support large-scale, population-wide screening. In fact, no such system exists today despite the potential impact. The most recent and most complete related work is the polyp detection system Polyp-Alert [62], which can provide near real-time feedback during colonoscopies. However, it is limited to polyp detection, and it is not fast enough in the case of live examinations. To detect mucosal lesions in the colon, we built a system combining filters using machine learning, image recognition, and extraction and comparison of global and local image features. Furthermore, it is easy to add new filters or other types of data, for example, patient records or sensor data, to increase accuracy or enable detection of other pathologies. As a first step, we evaluate our prototype by training classifiers that are based on the different image recognition approaches. It is important to point out that these classifiers can also process other input, for example, sensor data. One example of experiments we are performing is for longitudinal GI of the patient being screened, one where previous abnormality data is pulled from the patient's journal and aligned algorithmically with current abnormalities. The goal of this is the ability to visualize and capture the development of individual abnormalities over time.

We also test the generated classifiers with different data and thereby evaluate the different approaches for feasibility of colonic polyp recognition and localization. The initial 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 system performance reaches real time in terms of video processing up to high-definition resolutions, and finally, (iii) that our system is using an acceptable amount of resources regarding memory consumption and CPU. This latter property makes our system potentially scalable with more data and different diseases to detect in parallel at runtime. This is an important requirement if, as we plan next, to put it to real use in a more clinical context.

The rest of the article is organized as follows: In Section 2, we briefly introduce our medical case study. This is followed by a presentation of the complete system in Section 3. Subsequently, we present a detailed evaluation of the whole system in Section 4, and in Section 5, we discuss two cases where our system will be used in two medical examinations. We present related work in the field and compare it to the presented system in Section 6. Finally, we draw conclusions in Section 7.

2. GASTROINTESTINAL ENDOSCOPY

The complex GI system can be affected by various diseases; CRC is one of the major health issues world wide. Some examples of these diseases and their complexity can be seen in Figure 1. If CRC is detected at an early stage, the prognosis is substantially improved, from a 90% 5-year survival probability in the early stage 1 to only 5–10% 5-year survival probability in the latest stage 4 [5]. Several studies have shown that large population-based screening programs improve the prognosis and even reduce incidences of CRC [19], and the European Union guidelines recommend screening for CRC for all persons older than 50 years [56].

GI endoscopies are common medical examinations where the lumen and the mucosa of the entire GI tract are visualized to diagnose diseases [34]. The endoscopic system is made of an endoscope, a flexible tube with a charge couple device (CCD) chip and two bundles of optical fibers at the tip. The endoscope is connected to a video processor and a light source, and the video signals are transferred to a screen for the doctor to analyze. The common gold standard GI endoscopic examinations are gastroscopy and colonoscopy. However, such endoscopies are demanding and invasive procedures, and can be of great discomfort for patients. They are performed by medical experts (endoscopists), have to be performed in real time, and do not scale well to larger populations due to labor-intensive expert involvement. Additionally, the procedure is expensive. In the US, for example, colonoscopy is the most expensive cancer screening process with annual costs of 10 billion US dollars (USD 1,100/person) [55], with a time consumption of about 1 medical-doctor-hour and 2 nurse-hours, per examination. Furthermore, colonoscopy is not the ideal screening test; many polyps are hard to detect (Figure 1(f)), and in average, 20% of polyps are missed or incompletely removed, i.e., the risk of getting CRC later on largely depends on the endoscopist's ability to detect polyps [23]. We therefore aim for a system that detects mucosal pathologies in videos of the GI tract where the goal is to assist endoscopists during live examinations.

Once a polyp is detected, the morphology needs to be assessed to determine whether or not the polyp has a risk of malignant transformation. There exist mainly three classification systems for polyp assessment, two for characterization of the surface and one for the shape. The Kudo and the Nice-classification are both used to characterize the surface structure of the polyp. The Kudo-classification [27] is based upon chromoscopy requiring supplementary staining of the mucosa with a colorant, while the Nice-classification [38] is based on electronic color filter on the scope. The Paris classification is used to describe the shape of the polyp [21]. Despite these classifications, endoscopists assess polyps quite differently, and a standard computer algorithm for interpretation may therefore reduce the differences in the assessment [12].

Moreover, alternatives to traditional endoscopic examinations have recently emerged with the development of non-invasive VCEs. A pill-sized camera (available from vendors such as Given and Olympus) is swallowed and next records a video of the entire GI tract. The challenge in this context, at least if the examinations should be scaled to everyone above 50, is that endoscopists still need to analyze the videos. This creates an impractical scaling problem due to a limited number of endoscopists, which is one important motivation for developing our EIR system. Thus, in the VCE context, EIR is built for first-order, large-scale screening to determine whether a traditional endoscopic examination is needed or not, i.e., limiting and reducing the traditional endoscopy examinations to patients with positive findings from the VCE examination.

Consequently, we aim for a multimedia analysis system that can be used both as a live computer aided diagnostic system and as an automatic detection system for screening systems using VCEs. As a first step, we target detection of colorectal polyps (see, for example, Figure 1(a)). The reason for starting with this scenario is that most CRCs arise from benign, adenomatous polyps containing dysplastic cells, and detection and removal of such polyps prevents the development of cancer. Nevertheless, our system will be extended to support detection of multiple abnormalities and diseases of the GI tract by training the classifiers using different datasets.

3. EIR ARCHITECTURE

Based on the two target use-cases, the main objectives of the EIR system are (i) easy to use, (ii) easy to extend to different diseases, (iii) real-time handling of multimedia content, (iv) being able to be used as a live system, and (v) high classification performance with minimal false-negative classification results. It can be split into three main parts: the annotation sub-system, the detection and automatic analysis sub-system, and the

visualization and computer-aided diagnosis sub-system. All three parts are important to achieve a holistic system that can support doctors in disease detection and diagnosis in the GI tract.

3.1. Annotation Sub-system

The main purpose of the annotation sub-system is to collect training data for the detection and automatic analysis sub-system. This type of data can only be collected with the help of medical experts. To make the collection process easier for the doctors and as efficient as possible, we combine manual annotations with automatic methods. It is well known that training data is an important key factor to create a good classification system. Nevertheless, in the medical field, the number of available experts and the multimedia data are two resources that are quite limited. This is primarily because of a high every-day workload for doctors, but also due to legal issues. In many countries, patient consent has to be collected before images or videos can be used, making it a very cumbersome task. Moreover, the annotation of videos itself is very time-consuming, and the quality of annotations depends on the experience and concentration of the doctors [18]. For example, in a VCE procedure, depending on the time the capsule needs through the GI tract, there are, on average, about 216,000 images per examination, and an endoscopist frequently needs 60 minutes and even up to 2 hours to view and analyze all the video data [29]. Therefore, besides getting data for the EIR system to enable automatic screening, the annotation sub-system also makes it possible to use the annotated videos in a medical video archive for procedure documentation or teaching purposes. The current version of the annotation part consists of the semi-supervised annotation tool presented in the work of Albisser et al. [2] and the new cluster-based annotation tool.

Semi-supervised annotation tool. Using the semi-supervised tool [2], the doctors only have to provide annotations in a single frame of the video or image to reduce the time they need to spend on the whole process. The specialist's knowledge is ideally only required for the first very basic identification of abnormalities and to tag them accordingly. This manual step is done by selecting any regions of interest in a video or image sequence. The automatic step uses this information to track the regions of interest on previous and subsequent frames automatically. There is still a fair amount of manual work involved. However, using a suitable tracking algorithm substantially reduces the time needed to create a complete dataset. Moreover, a lot of annotation work can be performed without the specialist being present all the time. The output generated by the tool is a list of frames for a certain disease including rectangles for every previously marked region within the frame. This data is especially helpful for training and development of localization and tracking algorithms.

Cluster-based annotation tool. To extended the annotation tool, we implemented an extension that allows the doctors to utilize global features-based clustering to tag a large number of images in a short time. The clusters are created based on visual global image features that are also used in our classification sub-system, and the doctors can subsequently drag and drop images between different automatically created clusters and also annotate complete clusters. This application has two main advantages. First, it allows medical doctors to investigate and analyze vast collections of frames from endoscopic procedures by providing a configurable focus and context view based on frame similarity. Second, it grants for utilizing the focus and context view for annotation and tagging of the dataset, making it more accessible for complimentary information systems. The clustering annotation tool combines content-based similarity, unsupervised clustering (x-means), supervised clustering (k-means), and focus/context views. Figure 2 shows the interface of the clustering annotation tool. On the upper

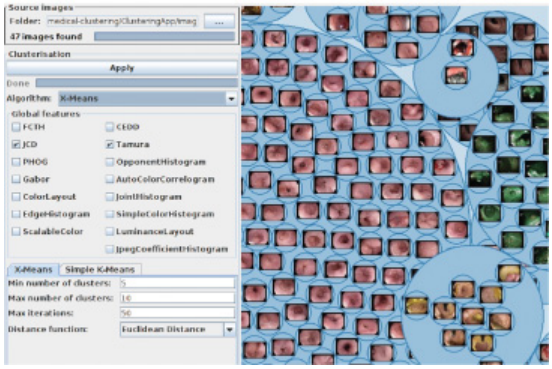


Fig. 2. Feature-based clustering annotation.

left side, users can choose the folder containing the image collection. The clustering algorithm can be selected in the setting below. For the clustering algorithm, several different image features can be chosen. If more than one feature is selected, they are combined using early fusion. The bottom options allow the user to specify the clustering parameters. These settings are set to default values recommended by literature. A click on the apply button creates the clusters and presents them on the right site. The cluster circles are represented using the image that is closest to the cluster center, i.e., the cluster medoid followed by the next closest and so on. The user can interact with the visual presentation by zooming and turning it into different angles. Furthermore, the user can double-click on clusters, which will open the folder containing all images in the selected cluster. The images can be dragged and dropped between different cluster circles, and with a right-click on the clusters, the user can see information like the cluster center and the purity of the cluster based on the distances. Finally, the medical experts can tag the clusters, which adds the tag to the name of the images in the cluster. The output of the clustering annotation tool is mainly used to identify and tag frames or images that contain abnormalities for the classification sub-system. Its output can also be used in the previous presented annotation tool to mark the exact position of abnormalities in the images.

3.2. Detection and Automatic Analysis Sub-system

Detection sub-system. The detection sub-system analyzes multimedia data, such as videos, images, and sensor measurements, to identify if there is anything abnormal to be found in the colon. All frames processed by this sub-system can be separated into two disjoint sets (positive and negative) which can also be seen as the model for the disease and abnormality detector. These two sets contain example images for abnormalities and images without any abnormality. The detection system is built in a modular way and can easily be extended with new models or sub-models. To compare and determine the abnormalities in a given video frame (or image), we use global image features, because they are easy and fast to calculate. We are not (yet) interested in the exact position for the detection sub-system. In previous work, we showed that global features indeed can outperform or at least reach the same results as local features [42]. EIR uses the Lire [32] open source library for content-based image retrieval. This library provides a comprehensive set of already implemented and tested algorithms to extract different types of global image features. This allows us to experiment with a whole set of global image features for detecting or clustering video frames from colonoscopy or VCE videos. Again, we do not claim novelty associated with individual algorithms and sub-components. Indeed, we carefully select and build on state-of-the-art technologies to get the optimal *integrated* holistic solution.

The indexing function is an extension of the indexing function used by Lire and provided by Lucene, modified with a hashing function which performs hashing on the given features and stores the hash values in the index. Lire uses Lucene inverted indexes for storing and searching image features data. Indexes are created using a merge-based data structure (k-way merge). The segments of the indexes are sorted in memory and then merged. Each newly added document (in our case, image) adds a new segment and is merged with the existing segments. This leads to average $b \times \log N$ indexes that are fast to update and also not too slow to search [17, 26, 48]. Furthermore, the structure of the index is field- and row-based where each row is defined by its fields. Example fields are image, binary values for the features or the hash value of the feature, and so on. The number of fields is variable depending on the number of used image features or metadata. The features are stored as a byte representation and as a text field containing hash values from a random projection hashing approach. The hashing is based on locality sensitive hashing (LSH). We use multiple random hash functions to hash the values of the features, which results in similar images getting the same hash values. Similar images are then hashed in the same hash bucket by a linear projection in random directions of the hash functions in the feature space of the image. Possible drawbacks of this method are that very ineffective hash codes can be created and a large number of hash tables is needed to achieve a reasonable search quality. Nevertheless, these drawbacks are acceptable compared to the increased speed of the search algorithm [50]. The used hash function $h(v) \in \{0, 1\}$ for a histogram v is defined as $h(v) = sgn(v \cdot r)$, whereas r is a random vector with evenly distributed elements $r_i \in [-w, w]$. n hash functions, then, are represented as one single hash value $H(v) < 2^n$ combined as a bit string. For indexing m hash values $H_j(v)$, $j \in [0, m\rangle$ hash values are generated. The used parameters for the hashing are $w = 2$, $n = 12$, and $m = 150$, which leads to a good tradeoff between search time and precision based on an evaluation of 100,000 test images.

The basic algorithm of our detection sub-system is based on an improved version of a search-based method for image classification presented in Riegler et al. [42]. The algorithm is basically a simple K-Nearest-Neighbor algorithm (k-NN). Normally, k-NN is a non-parametric algorithm, which means that the rank of the values are used rather than the parameters of each object. The classification is based on its k numbers of nearest neighbors by a majority decision. The differences to our used algorithm is that it is based on a ranked list of a search result, which is generated in real time for each query frame or image and that weighted values are used for finding a decision antithetical to the non-parametric-behavior of the standard k-NN. For the classification, three parts of a standard ranked search result list are used, i.e., the belonging class of each image in the list, the number of the occurrences of each class, and the position of the image in the ranked list as a weight. The algorithm is then defined as the following:

$$c = \arg\max_{c \in C} \left\{ ClassScore(c) = |c| \sum_{I_i \in \{I_i | Class(I_i) = c\}} \frac{1}{RankScore(I_i)} \right\}.$$

Class c is the class with the highest weighted $ClassScore$ of all classes $c \in C$, and $ClassScore$ is calculated by summing up the occurrences of each class c and multiplying it with the summed $WeightedRankScore$. $RankScore$ per class is calculated by dividing one by the rank for each search query. The $WeightedRankScore$ is the sum of all $RankScore$ in the rank list.

We create the indexes of as many example frames as we can get, but it is important to point out, as the experiments showed, that the detection indeed needs good training

data. However, the number of needed examples is rather low compared to other methods, for example, deep learning, which is known for its need for large and well-labeled datasets. The index also contains information about the presence and type of any disease in the frame or image. A classifier can then search the index for the frames that are most similar to a given input frame. Based on the classification of the results, the detection sub-system then decides which abnormality the input frame belongs to. The whole detection is realized with two separate tools, an indexer and a classifier. We have released the indexer and the classifier as a separate project called *OpenSea*.[1]

The purpose of the global image feature indexer is to extract visual features from input videos or images, and store these in the index. These indexes are used as input data for the search-based classifier. The indexer is created as a separate tool and in a way so that it is easy to distribute it over different nodes, using, for example, Apache Storm. The computational nature of the indexing part is similar to batch processing. Therefore, creating the models for the classifier could be done offline, and it is not influencing the real-time capability of the system because it is only done once at the very first time when the training data is inserted into the system. It creates indexes for all directories passed on from the system. The visual features to calculate and store in the indexes can be chosen based on the abnormality, because different types of diseases require different sets of features or combinations. For example, bleeding is easier to detect using color features, whereas polyps also require shape and texture information. The indexer processes all the frames in a given directory. It stores the generated indexes in a sub-directory inside the indexed directory. If multiple directories are passed for indexing, it creates a separate index for each directory. The classifier can be used to classify video frames from an input video into as many classes as the detection sub-systems model consists of. 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, and based on this, a decision is made. We refer to these previously generated indexes, which are searched for similar image features, as classifier indexes or indexes containing training data. The classifier expects at least one classifier index and an input source. The input source can either be a video, an image, or another previously generated index. The classifier is parallelized, and it can choose how many CPU cores to use or if GPUs should be used to improve the performance even more.

Localization sub-system. The detection sub-system cannot determine the location of the detected irregularity in a frame. This is the task of the localization sub-system which determines the exact position of the disease or abnormality (Figure 3). The localization sub-system analyzes video frames already marked to contain abnormalities by the detection sub-system, and these frames are then preprocessed by a sequence of various image processing procedures, resulting in a set of possible abnormality coordinates within each frame. Currently, the sub-system implements a model for polyp localization using a hand-crafted object localization method, based on the geometrical shape of polyps. The sub-system is written in C++, and it uses the OpenCV open source library for routine image contents manipulation and the CUDA framework for GPU computation support. The localization sub-system consists of two independent image processing pipelines: an image rectification and an abnormality localization pipeline. All the processed frames sequentially go through both pipelines. To evaluate the performance, both the image rectification and the polyp localization pipelines

[1]https://bitbucket.org/mpg_projects/opensea, released under GPLv3 (http://www.gnu.org/licenses/gpl-3.0.en.html).

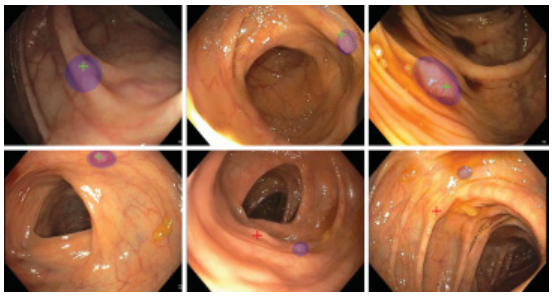


Fig. 3. The localization sub-system marks the possible locations of polyps. The first four show an exact match (ground truth marked with blue ellipses), but the last two are misses.

were implemented in two versions: a reference C++ code and a GPU-accelerated C++ code, with re-implementation of the most compute-intensive image processing steps as CUDA-kernels.

The image rectification pipeline uses pixel-level image processing in order to improve the overall image quality for the processing steps. Detected lesion objects 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, for example, seeds or stool, and lighted by direct light. Moreover, the image itself can be interleaved, noisy, blurry, and over-/under-exposed, and it can contain borders and sub-images. The images can also have various resolutions depending on the type of endoscopy equipment or VCE used. Endoscopic images usually have a lot of flares and flashes caused by high-power light sources located close to the camera. All these nuances negatively affect the local feature detection methods and have to be treated specially to reduce localization precision impact. In our case, we have used several sequentially applied filters to prepare raw input images for the following analysis by removing all the noisy artifacts. In particular, the current version of the system removes image borders, patients' data fields, imaging device state messages, embedded images, over- and under-exposed areas, and glare reflections.

The localization pipeline processes the rectified frames, and multiple pipelines for different abnormalities can run in parallel. The main idea of our localization algorithm is to use the polyps' physical shape to find the exact position in the frame. In most cases, the polyps have the shape of a hill located on relatively flat underlying surface or the shape of a more or less round rock connected to an underlying surface with a stalk varying in their thickness. These polyps can be approximated with an elliptical shape region that differs from the surrounding tissue. The polyp localization pipeline implements an image processing algorithm that performs, in sequence, the following steps: non-local means de-noising [6]; 2D Gaussian blur and 2D image gradient vectors extraction; border extraction by gradient vectors simple threshold binarization; removal of borders' isolated binary noise; possible location of ellipses focus estimation; ellipses size estimation by analyzing border pixels distribution; ellipses matching to extracted border pixels; selection of predefined number of non-overlapping local maximums and outputting their coordinates as possible polyp locations. For the possible locations of ellipses, we use the coordinates of local maximums in the insensitivity image, created by additive drawing of straight lines starting at each border pixel in the direction of its gradient vector. Ellipse matching is then performed using an ellipse fitting function [16].

All the constants and thresholds used in the image rectification and polyp localization algorithms are empirically selected from experimental studies and reflect nuances of the used data. They can be easily adjusted for different datasets, e.g., from another type of endoscope. The image rectification algorithm performs well for all medical imaging

artifacts lying outside the main image area. However, it should be improved to be able to detect and remove all pixels that belong to embedded images located anywhere in the frame. This is important for reducing the probability of false positive locations of findings inside of such embedded image regions. The polyp localization algorithm performs well for the used dataset and does not require training data for the detection. An example for the localization output with one possible polyp location is shown in Figure 3.

3.3. Visualization and Computer-Aided Diagnosis Sub-System

After the automatic detection and analysis of the content, the output has to be presented in a meaningful way to the medical expert. The visualization has to be reliable, robust, and easy to understand under stressful situations that can occur during a live examination. Furthermore, it has to support easy searches and browsing through large amounts of data. This is especially important for the VCE examinations due to the large amount of video material collected through such an examination (up to 12 hours). In general, the visualization sub-system has two main purposes. First, it should help in evaluating the performance of the system and get better insights into why things work well or not. Second, it can be used as a computer-aided diagnosis system for medical experts. In this context, we have the TagAndTrack tool [2] that can be used as a visualization and computer-aided diagnostic system. Furthermore, we developed a web technology-based visualization that is easy to use and distribute, and can be used to support medical experts during endoscopies. This tool simply takes the output of the detection and localization part and creates a web-based visualization, which is then combined with a video sharing platform where doctors are able to watch, archive, annotate, and share information. The information collected can later also be used for reinforcement learning in the detection and automatic analysis sub-systems.

4. SYSTEM EVALUATION

We have tested the system in terms of detection accuracy and system performance, and we also participated in a polyp detection challenge. All experiments are conducted on the same Linux machine with a dual 2.40GHz Intel Xeon CPUs (E5-2630), 16 physical CPU cores (32 with hyper-threading), 32GB of RAM, dual NVIDIA Corporation GM200 GeForce GTX TITAN X GPUs, a 256GB SSD and Ubuntu Linux. Furthermore, we used the ASU-Mayo Clinic polyp database as training and test data.[2] This dataset is the largest publicly available polyp dataset consisting of 20 videos, converted from WMV to MPEG-4 for the experiments, with a total number of 18,781 frames with $1{,}920 \times 1{,}080$ pixels resolution [52].

4.1. Detection Accuracy

For all detection and localization accuracy experiments, we used the common standard metrics precision, recall, and F1 score calculated on a per frame basis. This makes it more difficult for our algorithm to achieve good results, but it shows that the system works well. Furthermore, we decided to use leave-one-out cross-validation to evaluate this part of the system. Leave-one-out cross-validation is well-suited to show generalization potential and robustness of a predictive model. Therefore, the training and testing datasets are rotated, leaving out a single different non-overlapping video for testing, and using the remaining videos for training the model [13].

The developed system allows us to use several different global image features for the classification. The more image features we use, the more computationally expensive the classification becomes. Further, not all image features are equally important or provide

[2]http://polyp.grand-challenge.org/site/Polyp/AsuMayo/.

Table I. Leave-One-Out Cross-Evaluation Combined For All Supported Features

Feature	*True Pos.*	*True Neg.*	*False Pos.*	*False Neg.*	*Prec.*	*Recall*	*F1 score*
JointHist.	3,369	13,826	1,085	511	0.7563	0.8682	0.8084
JpegCoefficientHist.	3,224	13,772	1,139	656	0.7389	0.8309	0.7822
Tamura	3,392	13,861	1,050	488	0.7636	0.8742	0.8151
FuzzyOpponentHist.	3,341	13,552	1,359	539	0.7108	0.8610	0.7787
SimpleColorHist.	2,736	13,563	1,348	1,144	0.6699	0.7051	0.6870
JCD	3,556	13,777	1,134	324	0.7582	0.9164	0.8298
FuzzyColorHist.	2,708	13,243	1,668	1,172	0.6188	0.6979	0.6560
RotationInvariantLlBP	3,479	13,829	1,082	401	0.7627	0.8966	0.8243
FCTH	2,846	13,671	1,240	1,034	0.6965	0.7335	0.7145
LocalBinaryPatterns-AndOpponent	2,412	13,349	1,562	1,468	0.6069	0.6216	0.6142
PHOG	2,879	13,806	1,105	1,001	0.7226	0.7420	0.7321
RankAndOpponent	2,527	13,553	1,358	1,353	0.6504	0.6512	0.6508
ColorLayout	2,702	14,018	893	1,178	0.7515	0.6963	0.7229
CEDD	3,705	13796	1,115	175	0.7686	0.9548	0.8517
Gabor	1,849	10,643	4,268	2,031	0.3022	0.4765	0.3699
OpponentHist.	2,246	14,157	754	1,634	0.7486	0.5788	0.6529
EdgeHist.	3,548	13,737	1,174	332	0.7513	0.9144	0.8249
ScalableColor	3,231	13,684	1,227	649	0.7247	0.8327	0.7750
Late Fusion	3,710	13,894	1,017	170	0.7848	0.9561	0.8620

equally good results for our purpose. As a first step, we therefore need to determine which image features we want to use for classification. In order to understand which image features provide the best results, we generated indexes containing all possible image features for all frames of all video sequences from the test database. We can use these indexes for several different measurements and also for leave-one-out cross-validation. Using our detection system, the built-in metrics functionality can provide information on the performance of different image features for benchmarking. Further, it provides us with separate information for every single image feature, as well as the late fusion of all the selected image features. Moreover, literature indicates that late fusion approaches lead to a better performance than early fusion approaches [30, 49]. Escalante et al. [14], who came to the same conclusion, showed in their paper that late fusion performs well for multimedia retrieval tasks. They fused multiple heterogeneous image retrieval techniques developed for annotated collections. To perform late fusion, they used ranked lists created by search queries in their system to combine features. Based on the indication that late fusion is better suited for multimedia data, we use it for feature combination. Therefore, we classify each feature that we use separately, and combine them afterward using a majority decision weighted by the ranked score (an image class in a higher position in the ranked list gets a higher weight).

For our first experiment, we ran the detection with all possible image features selected, leaving out one video at the time, repeating the procedure until each video had been left out once. This is essentially the procedure for leave-one-out cross-validation. We then combined the reported values for true positives, true negatives, false positives, and false negatives for all the runs, and calculated the metrics for the combined values. The results of this first experiment are presented in Table I. All features used here are described in detail in the work of Lux [2013]. The single image feature that generally achieves the best score is Color and Edge Directivity Descriptor (CEDD). Further, the image features Joint Composite Descriptor (JCD), EdgeHistogram, Rotation Invariant Local Binary Patterns, Tamura, and Joint Histogram achieve promising results. The late fusion of all the image features achieves slightly better results. However, it is

Table II. Top 20 Feature Combinations Using Two Image Features for the Video wp_61, Sorted by F1 Score

Feature combinations	*True Pos.*	*True Neg.*	*False Pos.*	*False Neg.*	*Prec.*	*Recall*	*F1 score*
Rot.Inv.LBP/Tamura	162	22	153	0	0.5142	1	0.6792
PHOG/Tamura	161	23	152	1	0.5143	0.9938	0.6778
JpegCoeff.Hist./Tamura	162	21	154	0	0.5126	1	0.6778
Gabor/Tamura	162	20	155	0	0.5110	1	0.6764
FuzzyColorHist./Tamura	162	18	157	0	0.5078	1	0.6735
FuzzyOpp.Hist./FuzzyColorHist.	160	17	158	2	0.5031	0.9876	0.6666
JCD/Opp.Hist.	135	67	108	27	0.5555	0.8333	0.6666
JointHist./JpegCoeff.Hist.	162	12	163	0	0.4984	1	0.6652
ColorLayout /FuzzyColorHist.	162	11	164	0	0.4969	1	0.6639
FuzzyColorHist./JointHist.	162	11	164	0	0.4969	1	0.6639
FuzzyOpp.Hist./JointHist.	162	11	164	0	0.4969	1	0.6639
FuzzyOpp.Hist./SimpleColorHist.	162	11	164	0	0.4969	1	0.6639
JointHist./Rotat.Inv.LBP	162	11	164	0	0.4969	1	0.6639
JointHist./SimpleColorHist.	162	11	164	0	0.4969	1	0.6639
FuzzyOpp.Hist./Gabor	161	13	162	1	0.4984	0.9938	0.6639
JCD/JpegCoeff.Hist.	161	13	162	1	0.4984	0.9938	0.6639
CEDD/FuzzyColorHist.	159	17	158	3	0.5015	0.9814	0.6638
JpegCoeff.Hist./Rot.Inv.LBP	152	31	144	10	0.5135	0.9382	0.6637
JCD/Tamura	162	10	165	0	0.4954	1	0.6625
CEDD/Tamura	162	10	165	0	0.4954	1	0.6625

impractical to do a late fusion of all these image features as the calculation, indexing, and searching of all image features are computationally expensive. Therefore, we want to find a small sub-set of two image features, which provides optimal results despite minimizing the computational effort. Based on the evaluation results of different combinations of global features (Table II) using one video from the dataset, we decided that the image features JCD and Tamura seem to be the best combination for our performance measurements. The reason for this decision is because they have a good precision and recall, but at the same time, the computation time is low. We conducted this experiment only on one video to avoid optimizing our system on the used dataset, which could lead to results that do not really represent the true performance of the detection sub-system.

In these experiments, we also experienced that the only key parameter that influences the results in our classifier is the length of the ranked list. This has been set to 77 images based on the experiments because this is the value that gives a good tradeoff between precision and recall. A lower number of images in the ranked list leads to a higher precision, but a lower recall and vice versa.

To assess the actual performance of the classifier using these two image features, we again conducted a leave-one-out cross-validation with all available video sequences. The results are presented in Table III. With these settings, we achieve an average precision of 0.889, an average recall of 0.964, and an average F1 score value of 0.916. The problem with this average calculation is that different video sequences contribute values based on different numbers of video frames. If we weight the values contributed by every single video sequence with the number of frames in the sequence, we achieve an average precision of 0.9388, an average recall of 0.9850, and an average F1 score value of 0.9613. In other words, the results show that it is possible to detect polyps with a precision of almost 94%, and we detect almost 99% of all polyp containing frames. The results of these first experiments look very promising. Nevertheless, practical

Table III. Leave-One-Out Cross-Validation Using JCD and Tamura Features

Video	*True Pos.*	*True Neg.*	*False Pos.*	*False Neg.*	*Prec.*	*Recall*	*F1 score*
np_5	1	680	0	0	1	1	1
np_6	1	836	0	0	1	1	1
np_7	1	767	0	0	1	1	1
np_8	1	710	0	0	1	1	1
np_9	1	1,841	0	0	1	1	1
np_10	1	1,923	0	0	1	1	1
np_11	1	1,548	0	0	1	1	1
np_12	1	1,738	0	0	1	1	1
np_13	1	1,800	0	0	1	1	1
np_14	1	1,637	0	0	1	1	1
wp_2	140	9	20	70	0.875	0.6666	0.7567
wp_4	908	1	0	0	1	1	1
wp_24	310	68	127	12	0.7093	0.9627	0.8168
wp_49	421	12	62	4	0.8716	0.9905	0.9273
wp_52	688	101	284	31	0.7078	0.9568	0.8137
wp_61	162	10	165	0	0.4954	1	0.6625
wp_66	223	12	165	16	0.5747	0.9330	0.7113
wp_68	172	51	20	14	0.8958	0.9247	0.9100
wp_69	265	185	138	26	0.6575	0.9106	0.7636
wp_70	379	1	0	29	1	0.9289	0.9631
Average:					0.8890	0.9640	0.9160
Weighted average:					0.9388	0.9850	0.9613

suitability during live examinations comes with some difficulties. For example, during a live examination a lot of noise can occur, for example, instruments used, stool, and different lighting conditions. This is something that we want to explore in future work. To be able to do that, we collected a larger dataset that contains several different full-length procedures. We are currently working on the annotation of these videos. As soon as this is finished, more detailed and closer to real-world scenarios experiments will be conducted. We could also observe some variation in the precision and recall for some of the videos. A detailed investigation reveals that the detection part seems to be very accurate in detecting if a polyp is not there, but it is more difficult to find the correct frames that contain polyps based on the ground truth. Further investigations revealed that this is influenced by two aspects. First, because we use frame-based precision and recall, it is harder for the detection sub-system to achieve a high precision and recall. Second, because of the nature of the videos, the frames are often blurry (because of the motion blur), and it is hard to determine, even for a human observer, if the frame contains a polyp or not. A possible solution to solve this problem is to use time information of the videos to improve the classification performance, for example, by using the classification output of previous or next frames in the video to create an even more accurate classification output.

4.2. Localization Accuracy

We also used the common standard metrics precision, recall, and F1 score calculated on a per-frame basis for the localization accuracy experiments. It is important to point out that our localization algorithm does not require training like traditional learning-based algorithms. Therefore, all video segments were included in the experiments. As described previously, the localization sub-system is designed to process only frames that are marked to contain polyps by the detection sub-system. To evaluate the performance

Table IV. Performance of the Localization (Four Possible Polyp Locations Per Frame)

Data set	*True Pos.*	*False Pos.*	*False Neg.*	*Prec.*	*Recall*	*F1 score*
CVC-ClinicDB	397	215	249	0.6487	0.6146	0.6312
ASUMayo 2	1	244	244	0.0041	0.0041	0.0041
ASUMayo 4	443	467	467	0.4868	0.4868	0.4868
ASUMayo 24	74	300	300	0.1979	0.1979	0.1979
ASUMayo 49	36	355	355	0.0921	0.0921	0.0921
ASUMayo 52	194	490	490	0.2836	0.2836	0.2836
ASUMayo 61	129	80	80	0.6172	0.6172	0.6172
ASUMayo 66	92	142	142	0.3932	0.3932	0.3932
ASUMayo 68	63	126	126	0.3333	0.3333	0.3333
ASUMayo 69	0	235	235	0.0000	0.0000	0.0000
ASUMayo 70	4	381	381	0.0104	0.0104	0.0104
Average:				0.3207	0.3183	0.3195

of the localization system itself, we created a perfect-detection-dataset from the ASU-Mayo Clinic polyp database and the ground truth for polyp locations provided by it. The ground truth data is encoded as a set of images with the entire polyp area marked as a white pixel area on black background, one per original frame. A small amount of frames also contain more than one isolated polyp, which are counted as separate polyps. During the polyp location validation, we count each computed polyp location as true positive if the ground truth image has a pixel at the corresponding coordinates that is part of a polyp. Table IV presents the performance of the localization sub-system evaluation, with the output of four possible polyp locations per frame. The sub-system has a precision of 0.3207, a recall of 0.3183, and a F1 score of 0.3195. These results indicate that the localization part works as intended, but not perfectly. One reason that we identified for the sub-optimal performance of our algorithm is that it produces four possible disease locations per frame. Selection of multiple possible locations per frame is reasonable for the current localization sub-system version due to the lack of a tissue texture identification algorithm. It is not possible to distinguish between hill-shaped polyps and normal colon mucosa without corresponding textural analysis. Thus, multiple points finding increases the probability of hitting the polyp by, at least, one point out of four. For the evaluation, all points were included in the calculations, which influences the performance metrics negatively due to a high number of false positives. Regardless of the relatively low overall localization performance, the results of these first experiments look very promising. Nevertheless, the accuracy of the localization should be improved to make it suitable for practical use. We are currently working on an improved version of the algorithm that will include advanced shape and texture detection techniques together with inter-frame video sequence analysis.

4.3. MICCAI Challenge

To compare our method to other state-of-the-art methods, we participated in the Endovis Automatic Polyp Detection in Colonoscopy Grand Challenge[3] at the 2015 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). The challenge was divided into two parts. The first part was the polyp localization, where the question was whether the method could cope with important polyp appearance variability and, therefore, accurately determine the location of the polyp

[3]http://polyp.grand-challenge.org/.

Table V. MICCAI Polyp *localization* Challenge

Participant	True Pos.	False Pos.	False Neg.	Prec.	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	1,558	131	4.00	33.16	13.50
SNU	8	188	188	4.08	4.08	4.08

Table VI. MICCAI Polyp *Detection Latency* Challenge

Participant	Latency (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

in a frame. The second part was whether the method could detect a polyp in the frame or not, and how long the delay was from the first appearance of the polyp to when our system could detect it. In general, we did not expect very good results compared to the other specialized systems. Other participants used a wide range of different methods to detect polyps. These methods ranged from hand-crafted features, like contour or shape-based detection over machine learning approaches to neural networks. We identified several problem areas during the challenge such as blurry images due to camera motion, size differences, lighting, and objects that look like polyps, but are not, like contaminants.

Table V shows the result for the polyp localization part based on the CVC-ClinicDB dataset containing 612 still images from 29 different sequences. Our system is on the fourth place out of six. Details about the implementation of the first three methods are not available, but almost all of them used deep learning. Based on the fact that our system is not built for only polyp detection, the results are still very satisfactory. It is also important to point out that the first three participants were organizers of the challenge and involved in the dataset collection, and so on. Table VI shows the results of the detection latency part. For the latency, our system could perform second best of all participants. This is a very good result and a positive confirmation about the real-time performance compatibility of our system. The approach of UNS-UCLAN is not able to distinguish between a frame with or without a polyp. All in all, the results of the challenge are positive for a system that is designed to be extendible and refinable for different diseases. We showed that we can compete and outperform other state-of-the-art approaches, which are designed for the specific problem of the challenge, without applying any adaptations to our system.

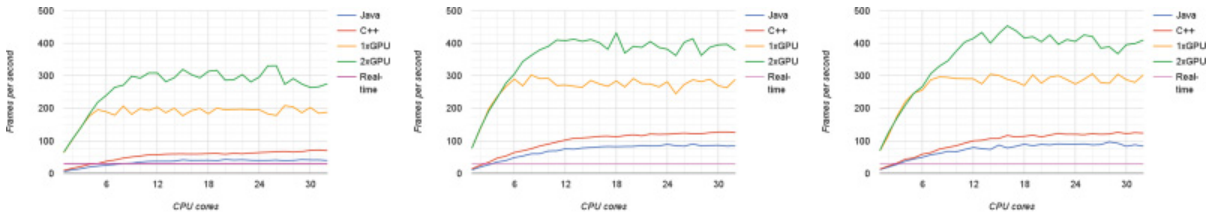
4.4. System Performance

A fundamental requirement of EIR is scalability and performance. The idea is to use the system for mass-screening for lesions in the GI tract, using video sequences recorded live with colonoscopy or VCEs, as well as a real-time diseases detection system that can be used during live endoscopy procedures. For the performance evaluation, we used the configuration of the system with best accuracy. This is rather obvious given our quest for a system that can be put to real use in clinical settings. Therefore, it is important to reach real-time performance in terms of processing a video and several other input signals at the same time and reach a frame rate of not less than 30 frames per second (FPS), which is the output of current endoscopes. For all the experiments, we used 20 videos from three different endoscopic devices and different resolutions, i.e., 1920×1080 (6 videos), 856×480 (4 videos), and 712×480 (10 videos).

4.4.1. Processing. To evaluate our detection sub-system, we first measured the indexing that creates the model later used by the classifier. This process does not need real-time performance and can be seen as batch processing, but it should at least be scalable for larger datasets.

Table VII. Indexing Performance of Four Different Datasets to Show the Scaling

Index	Frames	Total time in seconds	Time per frame in ms
$D1$	3,871	89.78	23.1
$D2$	14,909	178.55	11.9
$D3$	29,818	231.75	7.7
$D4$	100,000	782.351	7.8

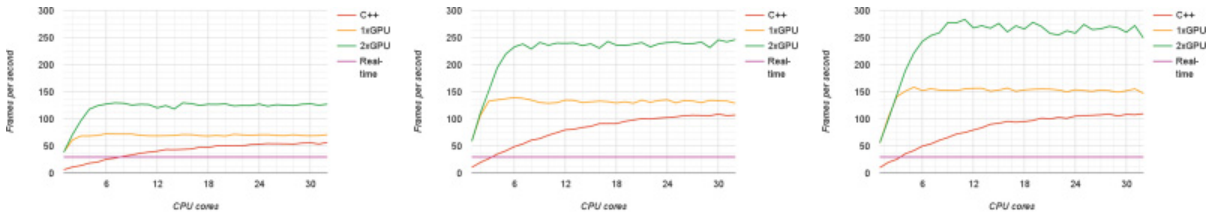


(a) Video resolutions of 1920 × 1080. (b) Video resolutions of 856 × 480. (c) Video resolutions of 712 × 480.

Fig. 4. The detection sub-system performance in terms of FPS depends on the number of CPU cores, the resolution of the videos, and the detection algorithm implementation.

Extracting two features and indexing them for the entire dataset take, on average, 5.2 milliseconds per frame. There is no big difference between the indexing time of different resolutions. We tested the scaling potential by indexing different datasets. The first dataset ($D1$) contains 3,871 frames, the second one ($D2$) contains 14,909 frames, the third one ($D3$) contains 29,818 frames, and the last one ($D4$) with 100,000 frames. Table VII shows the overall results. We discovered that a larger dataset leads to a faster indexing time per frame. We conjecture that this is due to reducing average per-frame processing overhead caused by GPU initialization and kernels loading into the GPUs. Furthermore, we did not find a significant increase after more than 30,000 frames in the dataset. The limiting factor is the I/O, since increasing the number of cores did not increase performance. All in all, our experiments show that the indexer is scalable in terms of larger datasets, and it should meet all requirements of the system for future tasks. The performance of the detection is also important, since the system should provide a result as fast as possible and not slower than 30FPS, making it usable for live applications. Again, we used the 20 different videos previously described. Figure 4(a) shows the detection sub-system performance in terms of FPS for the highest video resolution of 1920 × 1080. It depicts performance for all different detection algorithm implementations (Java, C++, and GPU) and different combinations of utilized hardware resources (from 1 to 32CPU cores and none, 1, or 2GPUs). For the full HD videos, the required frame rate of 30FPS is reached using 8, 5, and 1CPU cores in parallel for the Java, the C++, and the GPU implementations. Increasing the number of used CPU cores also increases the performance for all implementations, and the system reaches the maximum performance of 330FPS with 2GPUs and 25CPU cores. A slight decrease of the performance can be observed for a high number of used CPU cores. This is caused by an increased overhead for context switching and competition for resource. Figures 4(b) and (c) show the detection sub-system performance in terms of FPS for the videos with smaller resolution. The maximum performance of 430 (for 856 × 480 resolution) and 453 (for 712 × 480 resolution) FPS is reached using 2GPUs and 18 and 16CPU cores.

Figure 5(a) depicts the localization sub-system performance in terms of FPS for the highest quality video with a resolution of 1920 × 1080. Both the localization algorithm implementations (C++ and GPU) and different combinations of used hardware resources (from 1 to 32CPU cores and none, 1, or 2GPUs) are presented. For these videos,

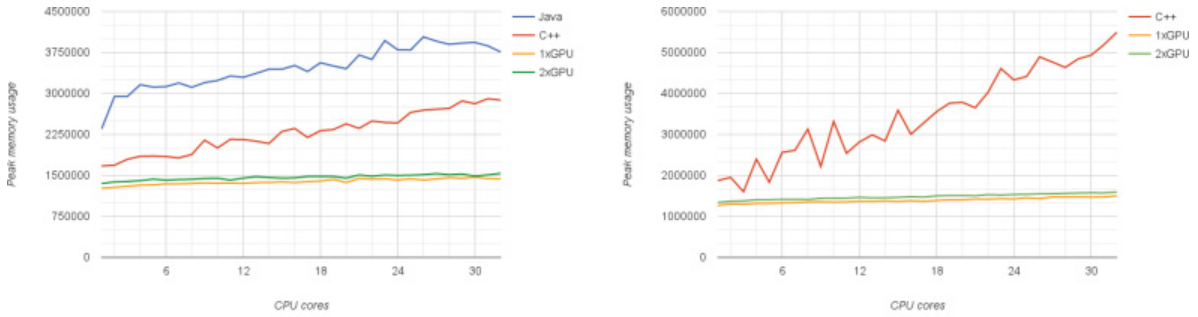


(a) Video resolutions of 1920 × 1080. (b) Video resolutions of 856 × 480. (c) Video resolutions of 712 × 480.

Fig. 5. The localization sub-system performance in terms of FPS depends on the number of CPU cores, the resolution of the videos, and the localization algorithm implementation.

the required frame rate of 30FPS is reached using 8 and 1CPU cores in parallel for the C++ and the GPU implementation. As expected, increasing the number of used CPU cores increases the FPS performance for both implementations and peaks at the maximum performance of 129FPS with 2GPUs and 15CPU cores. A slight decrease of the performance for a large number of used CPU cores caused by increasing overhead for context switching and resources competition happens also for the detection sub-system. Finally, Figures 5(b) and (c) show results for the videos with the smaller resolution. The peak performance of 246 (for 856 × 480 resolution) and 283 (for 856 × 480 resolution) FPS is reached using 2GPUs and 32 and 11CPU cores. The maximum GPU hardware utilization measured during our experimental studies was around 80% for both, using 1 or 2GPUs. The reason for the GPUs under-utilization is the implementation of some video frames processing algorithm steps on the CPU, namely the ellipse-shape detector, fuzzy logic for feature extractors, and building of frame features joint vector. This causes a large number of CPU-GPU data transfer and unavoidable GPU idling, required for the synchronization in multi-thread environments. Further implementations of other processing steps on heterogeneous architectures, such as GPUs, will lead to an increased performance and reduced utilization of the CPU resources. The outcome of these experiments clearly shows that our system can reach real-time requirements for the video processing and still has processing power left, which can be used to process other input data at the same time, for example, sensor data or patient records data. A number of complex features can be added into the detection and the localization sub-systems. This will increase the system's detection and localization accuracy and at the same time keep its ability to perform in real time. Moreover, it can also be used to process several data streams simultaneously in real time and significantly reduce the examination time of doctors. The time reduction lies around 5-10 times depending on the type of input data, like video resolution, framerate, and sensors used. Our evaluation also shows that this is a very complex topic and requires methods and technologies from several different multimedia research directions (signal processing, multimedia systems, information retrieval, deep learning, etc).

4.4.2. Data Handling. Figures 6(a) and (b) show the memory usage for both sub-systems. In the Java and the C++ implementations of the detection sub-system, as well as in the C++ implementation of the localization sub-system, the memory consumption behaves normally and shows that both sub-systems are scalable in terms of memory. The GPU implementations of both systems show an almost constant memory increase, which is caused by the used frame-by-frame processing scheme on the GPU devices. The results of the memory usage measurements for the various hardware configurations and video resolutions show that the maximum memory usage is less than 4.5GB for the detection and 6GB for the localization sub-system. This proves that the sub-systems consume a reasonable amount of memory, and therefore, memory is not a bottleneck for the scaling potential of the system.



(a) Detection sub-system. (b) Localisation sub-system.

Fig. 6. Overall memory consumption (resident set size).

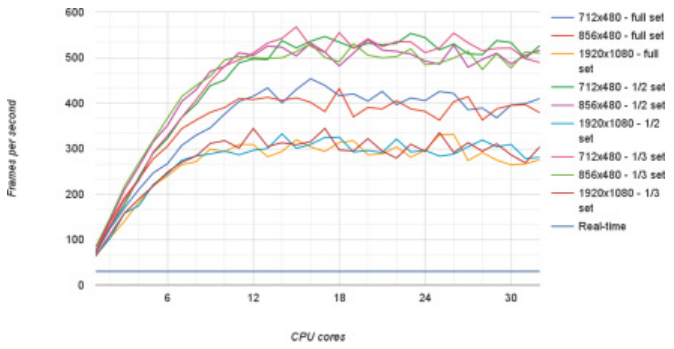


Fig. 7. Performance influence of different training data sizes for 1/2 and 1/3 of the original size.

A final question that we wanted to answer is if the size of the used classification indexes influences the detection accuracy or system performance. Figure 7 shows the system performance in terms of detection accuracy and FPS for three different training data sizes. The exception here was that smaller indexes would lead to a higher FPS throughput, but with a loss of classification performance. The experiment showed that the index size did not have a significant influence on the FPS output of the detection system. Another positive aspect is that the classification performance does not decrease with smaller indexes. The average F1 score for all three index sizes in this experiment increases with a decreasing index size. The index with the full training set reaches 0.938, the index that contains half of the training data (0.94) and the smallest index that only contains one third of the training data reaches an average F1 score of 0.946. This reveals that the detection sub-system also performs very well with a smaller amount of training data, which is a very positive point for the medical domain because of the lack of training data.

4.4.3. Distributed Processing Experiments. To investigate the performance on distributed hardware for the detection sub-system, some initial experiments on Amazon AWS EC2 instances were conducted. On a *c4.8xlarge* instance (Intel Xeon E5-2666 with 36 virtual CPUs), we were able to classify a video (MPEG-4) with 1,924 frames and a resolution of 1920 $\times$ 1080 using the JCD and Tamura features in 29.377 seconds with 65.5 FPS. When classifying data from a raw video file, the processing time increased to 39.599 seconds with 48.6FPS. When reading the data from a Windows media video (wmv) file, the processing time increased to 40.452 seconds with 47.6FPS. The *c4.8xlarge* instance is the most powerful instance offered by Amazon. Therefore, we conducted the same experiments on a less powerful *c4.4xlarge* instance (Intel Xeon E5-2666 with 16 virtual CPUs). Using this instance, we were able to process the MPEG-4 video data in 60.19 seconds with 31.97FPS, the wmv file in 81.17 seconds with 23.7FPS and the raw

video file in 79.718 seconds with 24.14FPS. This experiment shows that our system can be distributed, but using the given Amazon hardware, it did not really improve the performance when distributing the workload between several nodes. On the other hand, the performance using only local heterogeneous architectures easily meets the requirements, reducing the need for multi-machine distribution (for now).

5. REAL-WORLD USE-CASES

We are currently working on two different real-world use-cases with our partner hospitals. The first one is a live system intended to support and assist endoscopists while they perform live examinations. The second one has as a goal to automatically analyze videos captured by VCEs. The live system requires fast and reliable processing, and the VCE video analysis needs a system that is able to process a large amount of data fast, reliable, and in a scalable manner.

5.1. Live System

The live system is intended for the use-case where the endoscopist performs a routine examination. One screen shows the output of the colonoscope without the systems output. A second screen presents, in real time, the results of the algorithmic analysis to the doctor. In future clinical trials, we will evaluate and compare the current two-screen solution with a single screen combination. Previous studies have demonstrated that the detection rate of lesions is a major challenge [11, 54]. The aim of the live system is to use it as a visual recommendation toolkit for the human visual perception, much like a third, automatic eye with high-lighted sections to investigate/inspect more carefully by the doctor during the examination to improve the detection rate. While the endoscopist performs the colonoscopy, the system analyzes the video frames recorded by the colonoscope. At the beginning, we plan to show the physician optically (for example, with a red or green frame around the video) when the system detects a lesion in the actual frame or not. This can also be extended to the determination of what disease the system most probably detected and provide this information to the endoscopist. Apart from supporting the endoscopist during the colonoscopy, the system can also be used to document the procedure. After the colonoscopy, an overview can be given to the doctors where they can make changes or corrections, and add information. This can then be stored for later purposes or used as a written endoscopy report. Uninteresting parts of the video could be stored in a higher compressed way than important segments with the benefit of less storage space needed. Further, it would be practical to store high quality images of the most important parts. As de Lange et al. [11] show, single images can be an efficient way to store important findings from an examination. Another important part of computer aided live colonoscopies is the potential for temporal analysis when videos are captured multiple times from the same patient. Over the patient's medical history, analytics run on the same spatial colon parts to determine deltas (how development occurs) would be a meaningful addition to the now available standards and most probably improve the patients care and survival rate.

5.2. Wireless VCE

The multi-sensor VCE is swallowed in order to visualize the GI tract for subsequent detection and diagnosis of GI diseases. Thus, in the future, people may be able to buy VCEs at the pharmacy, and connect and deliver the video stream from the GI tract to the phone over a wireless network. The video footage can be processed in the phone or delivered to our system, which finally analyzes the video automatically. In the best case, the first screening results are available within 8 hours after swallowing the VCE, which is the time the camera typically spends traversing the GI tract. The

current VCEs have a low resolution of 256 × 256 with 3-30FPS (adaptive frame rate with a feedback loop from the receiver to the transmitter). They do not have optimum lighting, making it more challenging to analyze small details in the images. Nevertheless, ongoing work tries to improve the state-of-the-art technology, which will make it possible to use the methods and algorithms developed for colonoscopes also for VCEs [25]. In the case of the colon, accuracy of existing methods is far below the required precision and recall, and the processing of the algorithms does not scale in terms of high-volume data. Each type of disease or irregularity requires interaction between medical researchers dictating what the system must learn to detect, image processing researchers investigating detection or summarization algorithms, hardware developers to develop/produce/research sensors, and distributed processing researchers in order to scale the data analytics of the sensor data. For other scenarios, like in the upper part of the GI tract, there will be similar challenges and corresponding interaction between research disciplines. There are large challenges with respect to accuracy (precision and recall), scale of the processing, and hardware data quality because of different manufacturers (Olympus and Given are the market leaders). The aim is to be a major contributor in the area of medical imaging and sensor processing in the GI tract, as well as storing, processing, and analyzing this type of data.

6. RELATED WORK

A system aiming to analyze the whole GI tract needs to fulfill several requirements such as being able to process large amounts of data efficiently in real-time, while also being complete and practically applicable so that it can support doctors during colonoscopies or help analyzing data from VCEs. All requirements touch different areas of related work. In the following, we will discuss the most relevant works for our EIR system. Notice that no known existing complete algorithmic system is available, so we have to relate our work with others at the sub-component level.

Annotation. Liu et al. [31] describe a very advanced annotation tool called Arthemis. Arthemis is part of an integrated capturing and analysis system for colonoscopy, called Endoscopic Multimedia Information System (EMIS). EMIS provides functionality for collecting and archiving endoscopy videos. The use of an annotation tool for endoscopy videos is further researched by Lux and Riegler [33]. This demo paper focuses on common interaction methods for experts to annotate videos by recording speech and drawing onto the video. The paper aims at gathering information about the recorded videos in an easy and simple way, so that the annotation effort is minimally invasive for the daily routine of the experts. The related work in the field of annotation shows that it is crucial to integrate the annotation tool in a minimally invasive way within the environment of the experts. It is very important to provide them with a solution, which is very easy to use, and, at the same time, very easy to deploy in a restrictive medical environment. The annotation sub-system in EIR builds up on technologies and methods from the authors in Riegler et al. [43] and Lux [33].

Automatic analysis systems for the GI tract. Detection of diseases in the GI tract has mostly focused on polyps. This is most probably due to the lack of data in the medical field and polyps being a condition with at least some data available. Automatic analysis of polyps in colonoscopies has attracted research attention for a long time and several studies have been published [59, 60, 63]. However, there is a need for complete scalable real-time detection systems, both for computer aided diagnosis during colonoscopy examinations and for analysis of huge amounts of data from VCEs. Furthermore, all of the related works are limited to a very specific use-case, which in most cases is polyp detection for a specific type of camera. Several algorithms, methods,

Table VIII. Performance Comparison of Polyp Detection of State-of-the-Art Systems

Publication/System	What/Detection Types	Recall/Sensitivity	Precision	Specificity	Accuracy	FPS	Dataset Size
Wang et al. [62]	polyp/edge, texture	0.977*	–	0.957	–	10	1.8m frames
Tajbakhsh et al. [53]	polyp/shape, color, texture	0.5	–	–	–	–	35,000 frames
Park et al. [39]	polyp/shape, color, texture	0.828	0.658	–	–	–	62 images
Wang et al. [61]	polyp/shape, color, texture	0.814	–	–	–	0.14	1,513 images
Mamonov et al. [35]	polyp/shape	0.47	–	0.90	–	–	18,738 frames
Hwang et al. [20]	polyp/shape	0.96	0.83	–	–	15	8,621 frames
Li and Meng [28]	tumor/textural pattern	0.886	–	0.963	0.924	–	–
Zhou et al. [65]	polyp/intensity	0.75	–	0.959	0.908	–	–
Alexandre et al. [3]	polyp/color pattern	0.937	–	0.769	–	–	35 images
Kang et al. [24]	polyp/shape, color	–	–	–	–	1	–
Cheng et al. [7]	polyp/texture, color	0.862	–	–	–	0.08	74 images
Ameling et al. [4]	polyp/texture	AUC=0.95†	–	–	–	–	1,736 images
EIR	extendible/multiple	0.985%	0.939%	0.725	0.877	~ 75‡	18,781 frames

*The sensitivity is based on the number of detected polyps; other papers use per frame detection.
†Reported only area under the curve (AUC) instead of sensitivity.
‡Detection and localization performed together. Detection performance alone is around 300FPS and for localization around 100FPS.

and partial systems have been proposed and have achieved, at first glance, promising results in their respective testing environments. However, 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 dataset and test scenarios. Therefore, overfitting for the specific datasets can be a problem and can lead to unreliable results. Table VIII presents a summary of the most relevant approaches in colonoscopies and polyp detection. The last row of the table shows our approaches' performance to give a comparison. The first approach from Wang et al. [62] is the most recent and best working one in the field of polyp detection. A list of more related work can be found in their paper. As one can see in Table VIII, different methods provide different metrics for measuring the performance and use different datasets for training and testing. Moreover, almost all of them focus on polyp detection. Mamonov et al. [35] presented an algorithm for a binary classifier to detect polyps in the colon. The method is called binary classification with pre-selection, and it aims at reducing the amount of frames that need to be manually inspected. The sensitivity of the algorithm with regards to single input frames is significantly lower and only reaches 47%. A similar approach is presented by Hwang et al. [20]. This approach also focuses on shape, in particular on ellipses, which is a common shape for a polyp. Using this method, a frame is first segmented into regions by a watershed-based image segmentation algorithm. This algorithm is based on the observation that polyps are spherical or hemispherical geometric elevations on the surrounding mucosa. Similar to Mamonov et al. [35], they assume that multiple frames are available for one polyp and that a certain number of false negatives is acceptable in order to balance the number of false positives. The correctness of this assumption depends strongly on the frame rate of the camera that is used for recording the video. As mentioned in the introduction, the best working and complete system in the well-researched polyp detection field is Polyp-Alert [62], which is able to give near real-time feedback during colonoscopies. This approach is also listed as number one in Table VIII. The system can process 10 frames per second and uses visual features and a rule-based classifier to detect the edges of polyps. Further, they distinguish between clear frames and polyp frames in their detection. The researchers report a performance of 97.7% correctly detected polyps based on their dataset, which consists of 52 videos recorded using different colonoscopes. Unfortunately, the dataset

is not publicly available, and therefore, an exact detection performance comparison is not possible. Compared to our system, this system seems to reach higher detection accuracy, but it appears that our system is faster in terms of processing time per frame and can therefore detect polyps in real time. A comparison using the same hardware and full-length videos is currently to be carried out together with the developers of Polyp-Alert. Furthermore, our system is not designed and restricted to detect only polyps, and can be expanded to any possible disease if we have the correct training data. Another recent approach not limited to polyps is presented by Nawarathna et al. [36] describing a method to detect bleeding, but also polyps in colonoscopy videos.

Deep Learning. Deep learning is probably the most promising approach we need to explore further in EIR, and it is already very relevant for similar problems detecting, for instance, breast cancer [57], polyp detection [53], or lung cancer [9]. Nevertheless, such approaches are challenging to use in our use-case [8]. First, training is very complicated and time consuming. Our system has to be fast and understandable since we deal with patient data, where the outcome can differentiate between life and death. This can lead to serious problems in the medical field since it is very difficult to evaluate them properly [37]. Furthermore, one of the biggest challenges is that they require, most of the time, a lot of training data. In the medical field, this is a very important issue since it is hard to get data due to the lack of experts' time (doctors have a very high workload), and legal and ethical issues. Some common conditions, like colon polyps, may reach the required amount of training data for deep learning, 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 [46]. Nevertheless, for certain use-cases, such as presented in the work of Wang et al. [57], a small amount of training data can lead to reasonable results. As shown in Table VIII, recent neural network-based approaches for polyp detection are able to achieve interesting results, but still use relatively small labeled datasets in terms of the number of images or videos. Tajbakhsh et al. [53] presented a combined algorithm for a binary classifier to detect polyps in the colon, which was trained and tested on a 35,000 frames dataset with only 20 different polyps. The proposed polyp detection method first selects multiple possible polyp locations in a frame using machine learning of local polyp features such as color, texture, shape, and temporal information in multiple scales. A generated set of locations is then processed by a number of convolution feature-specialized neural networks and followed by results aggregation and frame binary classification. The detection performance of the method is 0.002 false positive per input frame at 50% sensitivity. A similar work is presented by Park et al. [39]. This approach focuses on shape detection via scale-invariant learning of hierarchical features using convolutional neural networks. Experimental results presented in the paper show that the method's sensitivity reaches around 83% with 66% precision on a 62 images dataset. Finally, it should be mentioned that neural networks are not easy to design for obtaining results that are explainable to a doctor. In a multi-class decision-based system, which is built to support medical doctors in decision-making, the fact why the system made certain decisions is important information. Approaches with a better understanding of the problem give a better explainable output that can be directly translated to the real-world scenario [51]. To test our assumptions about deep learning, we started conducting some experiments comparing deep learning approaches with our system. Initial experiments, based on implementations in Google Tensorflow [1] for the classification part and the YOLO [41] and Tensorbox[4] tracking algorithms for the localization part, revealed that our system can outperform or, at

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

least, reach the same single- and multi-class classification and detection performance as these systems, and that it is faster in training and new data processing if run on the same hardware configuration. We proved that our system can be easily extended adding new types of abnormalities. For the ASU-Mayo polyp dataset, the global feature approach reached a F1 score of 0.961 and the deep learning–based approach of 0.936. For our own created multi-disease dataset (which will be public available and shareable in the future), the global feature approach reached a F1 score of 0.909 compared to 0.875 for the deep learning approach. In the case of reduced amount of training data, our system seems to perform better, which is an important factor in the medical field. We conjecture that a combination of both approaches might be the best solution for future extensions of EIR, and detailed experiments are presented in the work of Pogorelov et al. [40].

7. CONCLUSION

In this article, a complete multimedia system for annotation, automatic disease detection, and visualization in context of the GI tract has been presented. Architecting the end-to-end EIR system has been largely motivated by the rapidly developing GI problems in the medical domain, combined with our bold idea that future GI screening can be performed relatively non-invasively at a scale where those interested can afford to be screened regularly, and it does nor require a quadrupling or so in number of GI specialists. An algorithmic end-to-end approach is a practical solution, and our EIR system is the first end-to-end multimedia GI system that is both accurate enough, and performs at a level where it can be used in real time. We described the whole system in detail from the annotation, automatic analysis, and detection to visualization. Further, we presented a detailed evaluation of the performance of the system in the area of detection accuracy, processing time, and scalability. The evaluation showed that the system achieves equal or better results than state-of-the-art in terms of accuracy, i.e., reaching a detection accuracy for polyps of more than 90% using the largest available dataset today (the ASU-Mayo clinic polyp dataset). On the other hand, our system outperforms other proposed systems when it comes to system performance. We showed that it is capable of scaling to fulfill big data requirements and that it can be used in real-time scenarios, i.e., in our live colonoscopy scenario, EIR processes HD resolution videos at about 300FPS. Moreover, we participated in a grand challenge to compare the system to other methods and could achieve good results for a very specific use-case with a system that is able to be used for many different use-cases at the same time. Additionally, we presented a real clinical setting implementation and use-case of our system that is currently being built with our hospital partners. For future work, we plan to include different abnormalities to detect and to even further improve the detection and localization accuracy. We are also collecting more training data and knowledge for the system with the help of medical experts from different collaborating hospitals in Sweden, Norway, Spain, Italy, and Japan. It is important to get data from different hospitals to be able to build a general system that is not shaped on a specific camera type or setup.

REFERENCES

[1] Martın Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A system for large-scale machine learning. *Proc. of USENIX OSDI*. 265–283.

[2] Zeno Albisser, Michael Riegler, Pål Halvorsen, Jiang Zhou, Carsten Griwodz, IIangko Balasingham, and Cathal Gurrin. 2015. Expert driven semi-supervised elucidation tool for medical endoscopic videos. In *Proc. of ACM MMSys*. 73–76.

[3] Luís A. Alexandre, Joao Casteleiro, and Nuno Nobreinst. 2007. Polyp detection in endoscopic video using SVMs. In *Proc. of PKDD*. 358–365.

[4] Stefan Ameling, Stephan Wirth, Dietrich Paulus, Gerard Lacey, and Fernando Vilarino. 2009. Texture-based polyp detection in colonoscopy. In *Bildverarbeitung für die Medizin*. Springer, 346–350.

[5] Hermann Brenner, Matthias Kloor, and Christian Peter Pox. 2016. Colorectal cancer. *The Lancet* (2016), 1490–1502.

[6] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. 2011. Non-local means denoising. *IPOL* 1 (2011), 208–212.

[7] Da-Chuan Cheng, Wen-Chien Ting, Yung-Fu Chen, Qin Pu, and Xiaoyi Jiang. 2008. 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*. Springer, 62–72.

[8] Christine Chin and David E. Brown. 2000. Learning in science: A comparison of deep and surface approaches. *Journal of Research in Science Teaching* 37, 2 (2000), 109–138.

[9] Francesco Ciompi, Kaman Chung, Sarah J. van Riel, Arnaud Arindra Adiyoso Setio, Paul K. Gerke, Colin Jacobs, Ernst Th. Scholten, Cornelia Schaefer-Prokop, Mathilde M. W. Wille, Alfonso Marchiano, and others. 2016. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *arXiv preprint arXiv:1610.09157* (2016).

[10] Yang Cong, Shuai Wang, Ji Liu, Jun Cao, Yunsheng Yang, and Jiebo Luo. 2015. Deep sparse feature selection for computer aided endoscopy diagnosis. *Pattern Recognition* 48, 3 (2015), 907–917.

[11] Thomas de Lange, Stig Larsen, and Lars Aabakken. 2005. Image documentation of endoscopic findings in ulcerative colitis: Photographs or video clips? *Gastrointestinal Endoscopy* 61, 6 (2005), 715–720.

[12] Ayso H. de Vries, Shandra Bipat, Evelien Dekker, Marjolein H. Liedenbaum, Jasper Florie, Paul Fockens, Roel van der Kraan, Elizabeth M. Mathus-Vliegen, Johannes B. Reitsma, Roel Truyen, and others. 2010. Polyp measurement based on CT colonography and colonoscopy: Variability and systematic differences. *European Radiology* 20, 6 (2010), 1404–1413.

[13] Bradley Efron and Robert Tibshirani. 1997. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association* 92, 438 (1997), 548–560.

[14] Hugo Jair Escalante, Carlos A. Hérnadez, Luis Enrique Sucar, and Manuel Montes. 2008. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proc. of ICMR*. 172–179.

[15] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. W. Coebergh, H. Comber, D. Forman, and F. Bray. 2013. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *European Journal of Cancer* 49, 6 (2013), 1374–1403.

[16] Andrew W. Fitzgibbon and Robert B. Fisher. 1995. A buyer's guide to conic fitting. *Proc. of (BMVC)*. 513–522. http://dl.acm.org/citation.cfm?id=243124.243148.

[17] The Apache Software Foundation. 2013. Apache Lucene - Index File Formats. Retrieved from https://lucene.apache.org/.

[18] B. Giritharan, Xiaohui Yuan, Jianguo Liu, B. Buckles, JungHwan Oh, and Shou Jiang Tang. 2008. Bleeding detection from capsule endoscopy videos. In *Proc. of EMBS*.

[19] O. Holme, M. Bretthauer, A. Fretheim, J. Odgaard-Jensen, and G. Hoff. 2013. Flexible sigmoidoscopy versus faecal occult blood testing for colorectal cancer screening in asymptomatic individuals. *The Cochrane Library*.

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

[21] H. Inoue, H. Kashida, S. Kudo, M. Sasako, T. Shimoda, H. Watanabe, S. Yoshida, M. Guelrud, C. J. Lightdale, K. Wang, and others. 2003. The Paris endoscopic classification of superficial neoplastic lesions: Esophagus, stomach, and colon: November 30 to December 1, 2002. *Gastrointest Endosc* 58, 6 Suppl (2003), S3–43.

[22] Menglin Jiang, Shaoting Zhang, Hongsheng Li, and Dimitris N. Metaxas. 2015. Computer-aided diagnosis of mammographic masses using scalable image retrieval. *IEEE Transactions on Biomedical Engineering* 62, 2 (2015), 783–792.

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

[24] J. Kang and R. Doraiswami. 2003. Real-time image processing system for endoscopic applications. In *Proc. of IEEE CCECE*.

[25] A. Khaleghi and I. Balasingham. 2015. Wireless communication link for capsule endoscope at 600 MHz. In *Proc. of IEEE EMBC*. 4081–4084.

[26] Donald Ervin Knuth. 1998. *The Art of Computer Programming: Sorting and Searching*. Vol. 3. Pearson Education.

[27] S. Kudo, S. Hirota, T. Nakajima, S. Hosobe, H. Kusaka, T. Kobayashi, M. Himori, and A. Yagyuu. 1994. Colorectal tumours and pit pattern. *Journal of Clinical Pathology* 47, 10 (1994), 880–885.

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

[29] Baopu Li and Max Q. H. Meng. 2009. Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments. *CBM* 39, 2 (2009), 141–147.

[30] Xirong Li, Cees G. M. Snoek, and Marcel Worring. 2010. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proc. of ACM ICMR*. 10–17.

[31] Danyu Liu, Yu Cao, Kihwan Kim, Sean Stanek, Bancha Doungratanaex-Chai, Kungen Lin, Wallapak Tavanapong, Johnny S. Wong, Jung-Hwan Oh, and Piet C. de Groen. 2007. Arthemis: Annotation software in an integrated capturing and analysis system for colonoscopy. *Computer Methods and Programs in Biomedicine* 88, 2 (2007), 152–163.

[32] Mathias Lux. 2013. LIRE: Open source image retrieval in Java. In *Proc. of the 21st ACM MM*. ACM, 843–846.

[33] Mathias Lux and Michael Riegler. 2013. Annotation of endoscopic videos on mobile devices: A bottom-up approach. In *Proc. of ACM MMSys'13*. ACM, 141–145.

[34] Shawn Mallery and Jacques Van Dam. 2000. Advances in diagnostic and therapeutic endoscopy. *Medical Clinics of North America* 84, 5 (2000), 1059–1083.

[35] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai. 2014. Automated polyp detection in colon capsule endoscopy. *IEEE Transactions on Medical Imaging* 33, 7 (2014), 1488–1502.

[36] Ruwan Nawarathna, JungHwan Oh, Jayantha Muthukudage, Wallapak Tavanapong, Johnny Wong, Piet C. De Groen, and Shou Jiang Tang. 2014. Abnormal image detection in endoscopy videos using a filter bank and local binary patterns. *NC* 144 (2014), 70–91.

[37] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2014. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv preprint arXiv:1412.1897* (2014).

[38] S. Oba, S. Tanaka, Y. Sano, S. Oka, and K. Chayama. 2011. Current status of narrow-band imaging magnifying colonoscopy for colorectal neoplasia in Japan. *Digestion* 83, 3 (2011), 167–172.

[39] Sungheon Park, Myunggi Lee, and Nojun Kwak. 2015. Polyp detection in colonoscopy videos using deeply-learned hierarchical features. *Proc. of (ISBI)*.

[40] Konstantin Pogorelov, Sigrun Losada, Carsten Griwodz, Thomas de Lange, Kristin Ranheim Randel, Duc Tien Dang Nguyen, Håkon Kvale Stensland, Francesco De Natale, Dag Johansen, Michael Riegler, and Pål Halvorsen. 2017. A holistic multimedia system for gastrointestinal tract disease detection. In *Proc. of MMSys*.

[41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2015. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640* (2015).

[42] Michael Riegler, Martha Larson, Mathias Lux, and Christoph Kofler. 2014. How 'how' reflects what's what: Content-based exploitation of how users frame social images. In *Proc. of ACM MM*. 397–406.

[43] Michael Riegler, Mathias Lux, Vincent Charvillat, Axel Carlier, Raynor Vliegendhart, and Martha Larson. 2014b. VideoJot: A multifunctional video annotation tool. In *Proc. of ACM ICMR*. 534–537.

[44] 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, and Pål Halvorsen. 2016. Multimedia and medicine: Teammates for better disease detection and survival. In *Proc. of ACM MM*. 968–977.

[45] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Thomas de Lange, Carsten Griwodz, Peter Thelin Schmidt, Sigrun Losada Eskeland, and Dag Johansen. 2016. EIR - efficient computer aided diagnosis framework for gastrointestinal endoscopies. In *Proc. of CBMI*.

[46] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61 (2015), 85–117.

[47] Joe V. Selby, Gary D. Friedman, Charles P. Quesenberry Jr, and Noel S. Weiss. 1992. A case–control study of screening sigmoidoscopy and mortality from colorectal cancer. *New England Journal of Medicine* 326, 10 (1992), 653–657.

[48] Theodoros Semertzidis, Dimitrios Rafailidis, Eleftherios Tiakas, Michael G. Strintzis, and Petros Daras. 2013. Multimedia indexing, search, and retrieval in large databases of social networks. In *Social Media Retrieval*. Springer, 43–63.

[49] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proc. of ACM MM*. 399–402.

[50] Jingkuan Song. 2013. Effective hashing for large-scale multimedia search. In *Proc. of Sigmod / PODS PhD Symp.* 55–60.

[51] Donald F. Specht. 1990. Probabilistic neural networks. *Neural Networks* 3, 1 (1990), 109–118.

[52] Nima Tajbakhsh, Suryakanth Gurudu, and Jianming Liang. 2016. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging* 35, 2 (Feb. 2016), 630–644.

[53] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. 2015. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. In *Proc. of IEEE ISBI*.

[54] Kyosuke Tanaka, Carlos A. Rubio, Aldona Dlugosz, Kotryna Truskaite, Ragnar Befrits, Greger Lindberg, and Peter T. Schmidt. 2013. Narrow-band imaging magnifying endoscopy in adult patients with eosinophilic esophagitis/esophageal eosinophilia and lymphocytic esophagitis. *Gastrointestinal Endoscopy* 78, 4 (2013), 659–664.

[55] The New York Times. 2013. The $2.7 Trillion Medical Bill. Retrieved from http://goo.gl/CuFyFJ.

[56] L. von Karsa, J. Patnick, and N. Segnan. 2012. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First edition–executive summary. *Endoscopy* 44 Suppl 3 (2012), SE1–8.

[57] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H. Beck. 2016b. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016).

[58] Shuai Wang, Yang Cong, Huijie Fan, Lianqing Liu, Xiaoqiu Li, Yunsheng Yang, Yandong Tang, Huaici Zhao, and Haibin Yu. 2016. Computer-aided endoscopic diagnosis without human-specific labeling. *Transactions on BME* 63, 11 (2016).

[59] Yi Wang, Wallapak Tavanapong, Johnny Wong, JungHwan Oh, and Piet C. de Groen. 2011. Computer-aided detection of retroflexion in colonoscopy. In *Proc. of IEEE CBMS*. 1–6.

[60] Yi Wang, Wallapak Tavanapong, Johnny Wong, JungHwan Oh, and Piet C. de Groen. 2013. Near real-time retroflexion detection in colonoscopy. *IEEE Journal of Biomedical and Health Informatics* 17, 1 (2013), 143–152.

[61] Yi Wang, Wallapak Tavanapong, Johnson Wong, JungHwan Oh, and Piet C. de Groen. 2014. Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy. *Journal of BMHI* 18, 4 (2014), 1379–1389.

[62] Yi Wang, Wallapak Tavanapong, Johnny Wong, Jung Hwan Oh, and Piet C. de Groen. 2015. Polyp-alert: Near real-time feedback during colonoscopy. *Computer Methods and Programs in Biomedicine* 120, 3 (2015), 164–179.

[63] Yi Wang, Wallapak Tavanapong, Johnny S. Wong, JungHwan Oh, and Piet C. de Groen. 2010. Detection of quality visualization of appendiceal orifices using local edge cross-section profile features and near pause detection. *IEEE Transactions on Biomedical Engineering* 57, 3 (2010), 685–695.

[64] World Health Organization - International Agency for Research on Cancer. 2012. Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. Retrieved from http://globocan.iarc.fr/Pages/fact_sheets_popula tion.aspx.

[65] Mingda Zhou, Guanqun Bao, Yishuang Geng, B. Alkandari, and Xiaoxi Li. 2014. Polyp detection and radius measurement in small intestine using video capsule endoscopy. In *Proc. of BMEI*. 237–241.

Received March 2016; revised March 2017; accepted April 2017

Paper VII

Multimedia and Medicine: Teammates for Better Disease Detection and Survival

Multimedia and Medicine: Teammates for Better Disease Detection and Survival

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[†§]

[†]Simula Research Laboratory, Norway [§]University of Oslo, Norway [⋆]Cancer Registry of Norway
[⊗]Klagenfurt University, Austria [◇]Vestre Viken Hospital Trust, Norway [◦]UiT - The Artic University of Norway
[⊕]Iowa State University, USA [▷]University of Catania, Italy [◁]Dublin City University, Ireland
[•]Karolinska Institute, Sweden [‡]Center for Digestive Diseases, Solna & Karolinska University Hospital, Sweden

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.

CCS Concepts

•Information systems → Multimedia information systems; •Applied computing → Health care information systems;

Keywords

Multimedia; Medical; Multimedia System

MM '16, October 15 - 19, 2016, Amsterdam, Netherlands

ISBN 978-1-4503-3603-1/16/10…$15.00

DOI: http://dx.doi.org/10.1145/2964284.2976760

1. INTRODUCTION

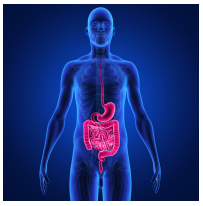


Figure 1: GI tract (shutterstock.com)

It is a typical assumption that visual analysis as it is already provided by the computer vision and medical image processing communities today is sufficient to solve health care multimedia challenges. Although we concede that computer vision and medical imaging methods are indeed essential contributors to promising approaches, we have come to the understanding that analyzing images and videos alone do not solve the challenges in medical fields such as endoscopy or ultrasound. Existing computer vision approaches do not make serious use of the multitude of additional information sources including sensors, temporal and users information.

Multimedia approaches are able to go beyond visual signals and also make use of heterogeneous sources including, e.g., the position sensors or fiber length measurement. Instead of considering the potential weakness of such signals as a nuisance, multimedia researchers are able to find ways to exploit them in combination to achieve the best possible results given the information available. Last but not least, multimedia cares first and foremost about the human user and assesses the feasibility of the resulting system. Correct and accurate diagnosis, efficient examinations and scalability are all critical for a health care system.

On the basis of these considerations, it is clear that we need to work on the challenge of realizing medical multimedia systems, which we define as follows: *a medical multimedia system is an interactive system, which provides support for diagnostics, examination, surgery, reporting and teaching in a medical setting by combining all available information sources and putting them in the hands of medical professionals or patients*. We note that some medical information systems may be fully automatic, but we still consider them to be at some level interactive, since a medical professional and/or a patient must be in the loop to interpret and act on the results.

In some areas of the human body, such as the gastrointestinal (GI) tract – our focus in this paper – the detection of abnormalities and diseases directly improves the chance of successful treatment, if the initial observation of disease indicators can be made visually, and also *before* the patient notices any symptoms. The GI tract is important since it is the site of many common diseases with high mortality rates. For example, three of the six most common cancer types are located in the GI tract (Figure 1), with a large number of

cancers detected yearly and with a high mortality rate [41]. Section 2 provides more details about diseases of the GI tract and their relevance, but clearly, early detection is important for patient survival. Currently, the recommended procedure for disease detection is gastrointestinal flexible endoscopy, i.e., the use of a flexible tube containing a lens system (cf. Figure 2(a).) Early detection and removal of cancer precursors to reduce cancer incidence makes regular screening of defined cohorts of the population necessary. Its implementation is obstructed by low willingness to undertake the unpleasant procedure, but also by inhibitive resource consumptions, and particular in terms of time required from the limited number of qualified medical staff. Alleviating these two limitations is essential and demands research into less intrusive detection procedures and an increased automatization of both detection and analysis of abnormalities.

There is a multitude of different use cases for automated diagnosis support, even within the limited field of GI tract inspection, which provide different opportunities beyond image analysis, and which require different kinds of assistance for medical experts. In our case, the use cases range from training support through archival, retrieval, and summarization for offline analysis to real-time annotation during endoscopy. The following quote from one of our discussions with medical specialists in endoscopy is bound to trigger the imagination of multimedia researchers with its hints for potential use cases:

"I am performing thousands of endoscopies, but I still miss abnormalities and have difficulties to analyze what I see. I would have liked more assisted examinations, and there is no possibilities to share these data with my colleagues or retrieve them when needed. It is just stored on a computer somewhere. I don't know where, and I don't think the IT support knows either... Sadly, we are collecting a lot of data, but we do not benefit from it at all. Do you have any idea what we can do with such data? I would be for example really nice if I could search for similar cases in our image database or use it to create automatic report. Reporting steals a lot of our time every day." – A Norwegian doctor, September 2015.

This quote directly reveals the need for real-time video analysis, storage, indexing, sharing and retrieval, audio transcripts, automatic annotation, action recognition, and probably more. After listening to this and many similar statements about insufficient time for manual analysis and unused multimedia data, we teamed up with specialists in the area of GI diseases to investigate how multimedia research can improve medical systems and patient treatment.

To aid and expand GI tract examinations, we have started the development of a multimedia system, which is called EIR after the Norse goddess of medical skills. It supports endoscopists in the detection and interpretation of diseases in the entire GI tract. Our aim is to develop both, (i) a live system assisting the detection and analysis of irregularities during colonoscopies and (ii) a future fully automated screening for the GI tract using a wireless video capsule endoscope (VCE).

In the **first use case**, we consider the provision of live assistance during classical colonoscopy. To support live colonoscopy while the procedure is running, the live-assisted system must process the input video stream from the endoscope (shown in Figure 2(a)) in real-time, and indicate automatically detected polyp candidates on a live video feed from the endoscope.

This approach is not meant to reduce the attention that medical doctors (endoscopists) performing a colonoscopy have to pay to the endoscopic video. It is rather meant to reduce the number of overlooked abnormalities and assist in the assessment of abnormalities, for example by providing size estimates and surface structure analysis to ease the distinction of polyps and regions that

(a) Colonoscopy equipment

(b) VCE capsule (camera pill)

Figure 2: Endoscopy vs. wireless capsule endoscopy (VCE).

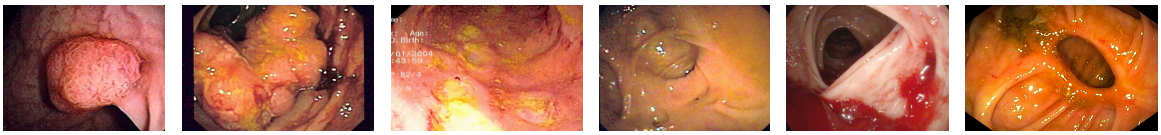
should raise concern from those that are better ignored. Obviously, live assistance has in the past been inhibited by excessive hardware costs, which prevented the creation and deployment of system that could perform in real-time. Our experimental prototype described in Section 4 makes use of modern parallel hardware, and shows very promising results, although we have only scratched the surface of the problem.

Our **second use case** is relevant in scaling GI tract examination to population-wide screening. This use case imposes strict requirements on the accuracy of the detection to avoid false negative findings (overlooking a disease). It is also challenging in terms of resource consumption, but the most precious resource in this case is the time required of endoscopists.

We believe that screening can become feasible through the use of VCEs (shown in Figure 2(b)), which can reduce several of the inconveniences and burdens of flexible endoscopy, although its current technical restrictions limit its usefulness. Nevertheless, while VCEs that could provide sufficient information were out of reach just a few years ago, it is now up to us to investigate the appropriate trade-off decisions on the recording side, which must consider frame rate, frame rate variability, scene lighting, storage space, resolution, quantization, energy consumption, detection rate and more. When we solve this challenge, VCEs become useful for the physician if the six to eight hours long video of the VCE's travel through the human GI tract can be summarized automatically in less than an hour. Such summarization is dominated by the challenges of unsupervised recording and the subsequent need to avoid false negatives.

We hope that our paper encourages the multimedia community to help improving the health care system by applying their knowledge and methods to reach the next level of computer and multimedia assisted diagnosis, detection and interpretation of abnormalities. In this area, computer vision and medical imaging have created visual representations of the interior of a body. To automatically detect and locate abnormalities, visual representations are not sufficient. There is a need for image and video processing, analysis, information search and retrieval, combination with other sensor data, assistance by medical experts, etc. – clearly multimedia – and it all needs integration and efficient processing. Therefore, in this paper, we look beyond computer vision and medical imaging and show the potential of multimedia research and that it goes far beyond well-known scenarios like analysis of content on YouTube and Flickr.

The paper is structured as follows. First we give an overview of health care multimedia challenges focusing on the field of GI endoscopy as an example of a medical field. That is followed by an overview of related work and current technologies. After that we present a showcase for a multimedia system for GI endoscopies to discuss the complexity and possibilities of medicine teamed up with multimedia. This part is underlined by a preliminary results section that should give an idea how such a multimedia application can be evaluated and what is important. Finally and most important we give an outlook and a summary including detailed description of how multimedia can be applied and what is needed.



(a) Colon polyp (b) Colorectal cancer (c) Crohn's disease (d) Diverticulosis (e) Bleeding (f) Anastomosis

Figure 3: A non-exhaustive set of examples of abnormalities that can be diagnosed using colonoscopy.

2. HEALTH MULTIMEDIA CHALLENGES

There are large societal challenges in the health care systems worldwide. If we look at our GI tract case study, about 2.8 millions of new luminal GI cancers (esophagus, stomach, colorectal) are detected yearly in the world, and the mortality is about 65% [41]. In addition to these cancers, numerous other chronic diseases (see Figure 3) 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 have a significant impact on the patients' health-related quality of life [7] and gastroenterology is one of the largest medical branches.

Nevertheless, there are unmet needs and potentials for improvements, which can be remedied by introducing better and more efficient digital medical systems. For 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% [6].

The current European Union guidelines therefore recommend screening for CRC [36]. Several screening methods exist, e.g., fecal immunochemical tests (FITs), sigmoidoscopy screening, computed tomography (CT) scans and colonoscopy. However, in randomized trials, only endoscopic methods have shown a reduced CRC incidence. However, it is not the ideal screening test, for a number of reasons. Each examination demands a significant amount of time from a medical professional and the procedure is unpleasant and can cause great discomfort for the patient [35] (Figure 2(a)). Moreover, on average, 20% of polyps, precursors of CRC, are missed or incompletely removed, i.e., the risk of getting CRC depends largely on the endoscopist's ability to detect polyps [15].

Furthermore, there are high costs related to these procedures. In the US, colonoscopy is the most expensive cancer screening process with an annual cost of $10 billion dollars, i.e., an average of $1,100 per examination (up to $6,000 in New York) [32, 33]. In the United Kingdom, the costs are around $2,700 per examination [29]. To meet the need for cost-effectiveness, improved diagnostics and enhanced efficiency in health care systems, 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 a VCE (Figure 2(b)), the aim is to make these examinations more easily accessible for patients and participants in screening programs, i.e., making the public health care system more scalable and cost-effective. It is also important that multimedia researchers address some of the challenges identified in the EU health policy, implemented through the Health Strategy, specially in the topics of prevention, health care access equalization, maintaining health into old age, and dynamic health systems incorporating new technologies. The optimal goal is to contribute in the area of medical multimedia for analysis as well as storage and processing of this type of data. Such next-generation big data applications, especially in the area of medicine, are frontiers for innovation, competition and productivity [20], where there are large initiatives both in the EU [1] and the US [21, 2].

3. RELATED WORK AND NEW TRENDS

To the best of our knowledge, currently, no start-to-end interactive medical multimedia system for annotating and analyzing data and computer aided diagnosis for the medical field exists. If one takes a closer look into the work of computer vision or medical image processing, it becomes clear that the complete loop is not their main research interest. A complete medical multimedia system including different multimedia applications that can fulfill the visions and objectives of the medical field must (i) have high detection accuracy (sensitivity, recall, precision), (ii) have an extensible and adaptable processing pipeline, (iv) support real-time processing to provide live feedback during for example endoscopy examinations, (v) support large-scale batch processing of, for example, VCE videos, (vi) be privacy-preserving, and (vii) visualize detection feedback to medical personnel. Several generally relevant systems fulfilling parts of the requirement list exist, but very few target medical scenarios, and no existing multimedia system matches all these requirements.

3.1 GI Tract Endoscopy Technology

There are several providers of endoscopy systems and VCE devices. Last generation equipment for manual procedures like colonoscopy and gastroscopy provides video with high resolution and high frame rates. There is, however, no computer-aided diagnostic feedback. In this respect, Polyp-Alert [40] is the most promising with polyp detection capabilities, but with the main purpose of evaluating how well the procedures are performed. For live analysis of endoscopy videos, our target system aims to go far beyond the currently existing systems. The other approach to record videos of the GI tract is VCEs using a small capsule type device (a 11mm×25mm pill), which has at least one image sensor, antenna, battery, light source and wireless transceiver. The capsule is swallowed to record the GI tract. There are several vendors providing such capsules, like IntroMedic, CapsoVision, Medtronic (Given) and Olympus. The current VCEs often have a variable framerate (increasing the framerate to about 30-35 FPS when entering the small intestine), but a rather low resolution ranging from 256×256 to 400×600. One of the main challenges for use of VCEs is man-hours of medical staff required for analysis. There are about 216,000 images per examination, and a very experienced endoscopist needs at least 30 to 60 minutes to process the video and possible sensor data. Therefore, it is important to develop automatic methods that can reduce the burden on medical staff and speed up the analysis of the videos. Currently, the software can segment the videos and can allow endoscopists to fast forward and look at multiple videos at the same time (probably affecting the detection accuracy). Moreover, some software includes small detection components that provides only vague "hints", for example about the detection of the color red, which may indicate bleeding. Other main limitations with VCEs are that the lack of means for

cleaning particles (food/stool) in the bowels, and their uncontrolled forward movement through the bowel that cannot be guided to take a close-up picture or a tissue sample from detected lesions.

Compared to traditional endoscopy examinations, with VCE, patient discomfort is decreased, and the size of the examined cohort may be increased. However, the analysis still requires a huge amount of manual labor and the image quality is substantially lower. Our research targets a system providing a far more advanced computer-assisted disease detection in general, detecting endoscopic findings with high accuracy, with reduced compute-resource consumption, to increase the number of screened people without spending huge amounts of time on manual analysis.

Current systems use mainly video and images for analysis. However, there is a large potential for adding more information. For example, knowing the position of the camera (either VCE or endoscope may narrow down the search for endoscopic findings). Furthermore, the VCEs and endoscopes will in the future be equipped with new sensors for biomarkers (bacteria, DNA, RNA...) and pH-meters (acid) [12], and research introduces the idea of VCEs with "legs" for controlled movement and "arms" for taking samples and injecting medication locally [34].

3.2 Abnormality Detection

As described above, we target detection of abnormalities in the entire GI tract. Currently, most existing systems mainly aim for detection of polyps in the colon. The main reason is the high clinical relevance and prevalence of CRC. Several studies have been published, e.g., [10, 11, 14, 19, 22, 23, 24, 25, 37, 38]. These related papers address polyp detection in several different ways. For example by using neural networks or handcrafted features like detection of round or ellipse shapes [14, 19], and by detecting the circular content areas [22, 23]. In Table 1, we compare the most promising and relevant systems according to reported performance (though not tested on the same dataset, and not all report the same metrics). The most recent and complete system for polyp detection is Polyp-Alert [40], which is able to give near real-time feedback during colonoscopies (10 FPS) with a very high accuracy. However, not many complete multimedia systems exist, and none of them is able to do real-time detection for use as a live support system during procedures. This means that endoscopists have to re-visit the videos after procedures, adding to the typically already crowded schedule of medical experts. Furthermore, all of them are limited to a very specific use case, and they all fail in one or more of the requirements of a future automatic system. Thus, there are a lot of open challenges that can be addressed by the multimedia community. With EIR, as a first step, we already perform at the level of state-of-the-art systems (last row of Table 1). Our ambitions are (i) to extend and improve our prototype far beyond both the current version of EIR and state-of-the-art, but more importantly, (ii) to inspire other multimedia researchers to explore the medical field.

4. SHOWCASE FOR HOW-TO MULTIMEDIA IN MEDICINE

To show how complex the medical field is and why multimedia research is needed, we developed the EIR multimedia system for automatic disease detection in the GI tract. We target the entire GI tract because not just the colon (the focus of most of the computer vision and medical image processing community) can contain diseases that should be detected. Figure 4 gives an overview of this system. The main requirements of such a system are (i) ease of use, (ii) ease of extending to different diseases, (iii) efficient real-time handling of multimedia content for both scale (VCEs) and support

Publication/ System	What/ Detection Types	Recall/ Sensitivity	Precision	Specificity	Accuracy	FPS	Dataset Size
Wang et al. [40]	polyp/edge, texture	97.7%*	–	95.7%	–	10	1.8m frames
Wang et al. [39]	polyp/shape,color,texture	81.4%	–	–	–	0.14	1,513 images
Mamonov et al. [19]	polyp/shape	47%	–	90%	–	–	18,738 frames
Hwang et al. [14]	polyp/shape	96%	83%	–	–	15	8,621 frames
Li and Meng [17]	tumor/textural pattern	88.6%	–	96.3%	92.4%	–	–
Zhou et al. [42]	polyp/intensity	75%	–	95.92%	90.8%	–	–
Alexandre et al. [4]	polyp/color pattern	93.7%	–	76.9%	–	–	35 images
Kang et al. [16]	polyp/shape,color	–	–	–	–	1	–
Cheng et al. [9]	polyp/texture,color	86.2%	–	–	–	0.08	74 images
Ameling et al. [5]	polyp/texture	AUC=95%	–	–	–	–	1,736 images
EIR	extendible/multiple	98.5%	93.88%	72.5%	87.7%	∼300	18,781 frames

* The sensitivity is based on the number of detected polyps, other papers use per frame detection.

Table 1: Performance comparison of polyp detection approaches of state-of-the-art systems. Not all performance measurements are available ("–").

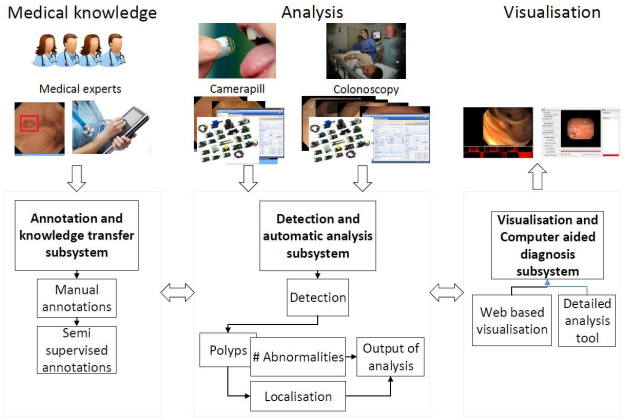


Figure 4: EIR system: annotation and knowledge transfer, detection and automatic analysis and computer aided diagnosis.

for live examinations, and (iv) high classification performance with minimal false negative classification results. To satisfy these requirements, the system has three main parts: The annotation and knowledge transfer sub-system, the detection and automatic analysis sub-system, and the visualization and computer aided diagnosis sub-system.

4.1 Annotation and Knowledge Transfer

The purpose of the annotation and knowledge transfer sub-system is to efficiently collect training data for the detection and automatic analysis sub-system. It is well known that training data is very important to make a good classification system. Additionally, in the medical field, the time of experts and annotated data are two very scarce resources. This is primarily because of high every-day workload for physicians, but also due to medical-legal issues. In terms of colonoscopy videos, the objective would be training a classifier for automatically detecting CRC, or its precursor lesions, colorectal polyps in multimedia data such as videos, sensor data and images. In our example system, we therefore developed an efficient semi-automatic annotation and knowledge transfer sub-system [3]. With a focus on ease of use and the minimal time requirements for annotation, our prototype was designed with a minimal level of required interaction.

The specialist's knowledge is only needed for the first identification of abnormalities and to tag them accordingly. This step is done manually by selecting any regions of interest in a video or image sequence and by annotation, i.e., providing information about importance and indicators for sensor data and patient records. After the manual annotation our prototype application uses object tracking to suggest annotations in further video frames by adjusting po-

sition and size of regions of interest as well as by automatically extending the annotation throughout a videos timeline. This data is then used in the analysis and detection sub-system. What we also have to learn from the medical doctors is how to interpret the various different data input sources, e.g., how to interpret the sensor data in the future, the significance of different pH (acidity) or biomarkers. It is important that multimedia researchers work hand in hand with the medical experts to gain this knowledge. Without efficient data collection tools, this will be an impossible task because of the time restrictions of medical personnel.

4.2 Detection and Automatic Analysis

The sub-system for detection and automatic analysis is designed in a modular way, making it possible to easily extend it to support different disease detectors, as well as other tasks like size determination and recognition of anatomical landmarks. Currently, it consists of two parts: (i) the detection sub-system that detects irregularities in video frames and images, and (ii) the localization sub-system that localizes the exact position of an abnormality in the frame. This part of the system is designed to detect whether there is something abnormal in a frame of the video (or image) or not. All the data that we process can be separated into two disjoint sets. These two sets contain example images, sensor data (temperature, blood, etc.) and other information that is useful for endoscopic findings, and images without any abnormality. It is important to point out, that the content based information images must be extended with other data like sensor output or information extracted from patient records to reach optimal results which makes it not a pure computer vision task. Each of these sets can be seen as the model for a specific disease. The modularity makes it possible to create a pipeline to for example first detect a polyp and then distinguish between a polyp with low or high risk of becoming a CRC by using for example the NICE classification[1]. To compare and determine the endoscopic findings in a given video frame, we use as a first approach global image features, i.e., because they are easy and fast to calculate, and at this stage, we do not need the exact position.

The basic idea is based on an improved version of a search-based method for image classification [27]. We chose this method because it is easy to implement and understand, and it gives us a first insight of the problem. Our experiments show that the detection needs good training data. However, the number of examples needed is rather low compared to other methods like deep learning. This is an important advantage at this point since there is not much data available. The classifier[2] tries to identify the frames that most probably contain a certain abnormality. Based on the classification of the results, the detection sub-system decides which endoscopic finding the input frame belongs to. This is done using late fusion of different classifiers. At the moment, we have one classifier for each global image feature. It is important to point out that the system will be expanded with other classifiers for sensor and audio data.

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 Lucene 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. Lucene indexes can contain all the information for one data point in one record (global features, sensor data, patient data, etc.). The system also includes a benchmarking function that will output evaluation information, and an HTML page with a visual representation of the results. For all video frames, we also can perform a localization. This is a pure computer vision problem and therefore we will not go in detail. It uses the information from the detection sub-system as a starting point, which means that it only processes frames that are already classified to contain an endoscopic finding. The processing of the images is implemented as a sequence of intra-frame pre- and main-filters. The output of this system can then further be used in for example a computer aided diagnosis program to help the doctor determining the size of a polyp or for reporting purposes.

4.3 Visualization and Diagnosis

One of the critical parts of each examination is the process of analyzing, reporting, facilitating and using multimedia to prepare the final result, i.e., the diagnosis and the report on the procedure. Medical doctors invest a significant part of their time on this task, and they are therefore in need of multimedia systems that help minimizing errors and increase the efficiency in this process.

For our experiments, we developed a web based visualization and annotation application to support medical experts with the goal of creating software that is easy to use and where it is easy to share data amongst participating medical experts. Our prototype facilitates the output of systems detection and localization part and creates a web based visualization which will be combined with a video sharing platform [13] where doctors are able to watch, archive, annotate and share information. We chose to use a centralized system based on web technologies to (i) minimize the necessary installs on client computers (with the current approach, a modern web browser is the only requirement), (ii) to allow for comfortable sharing of results and content with other experts, and (iii) to not duplicate data but use a centralized storage for multimedia data and annotations. This of course opens up questions about serving sensitive patient data over IP networks and leads to interesting research and organizational questions how to solve the data security problem, which is also an emerging field for the multimedia community, but data security is for now beyond the scope of the first EIR prototype.

While our first prototype is working as intended, the interplay between manually created content and automatically created content can still be improved. For example, applying object tracking algorithms is very difficult and often requires manual corrections. Most of the work in this step is done by the software end-users still need to navigate to the previously marked irregularities and playback the video from that point for the software to track the marked region on subsequent frames. Depending on the quality of the video and the speed of camera movement, user intervention is needed to assure a high quality of tracking. One can see, that there is still a fair amount of manual work involved, which makes it not really useful for medical experts. However, using a specialized – yet to be improved – tracking algorithm substantially reduces the time needed to, for example, create training videos or even datasets. Moreover, medical expert skills are maybe no longer necessarily required as the task of annotation correction is about tracking regions and adjusting rectangular dimensions rather than actually detecting or recognizing irregularities. This task could for example be outsourced using crowdsourcing. Our prototype visualization and annotation tool might be considered very basic, and there are tools resulting from multimedia research in existence that can be utilized for being a computer aided diagnosis system, but our approach already led to a benefit for the medical experts, allowing them to annotate and share data with other experts. Another area of multimedia, namely text-to-speech and text processing, could lead to great improvements in the reporting. When the endoscopic examination is completed the doctors have to transcribe what they visually observed into a written report following a standard proto-

[1] http://www.wipo.int/classifications/nice/en/

[2] To invite others to the area, we have released the basic algorithm as open source: *OpenSea*: https://bitbucket.org/mpg_projects/opensea.

col and using an internationally defined minimal standard terminology. This is a time consuming task and important information is sometimes forgotten or omitted. Consequently, computer based automatic transcription of audio information and combination of it with visual information in to a written patient record will probably increase the quality of the report and would substantially reduce the doctors workload. This will also make it possible to translate difficult medical terms into a report for the patient. Finally, not just the applications are important but also an understanding of how humans perceive multimedia content and how different aspects of the content influence them differently.

5. PRELIMINARY RESULTS

If multimedia researchers decide to work in the field of medicine we also have to make sure that our systems and applications are useful and accurate enough and achieve the required performance. Therefore, we tested our preliminary prototype in terms of accuracy and system performance. We used a computer with a dual 2.40GHz Intel Xeon CPUs (E5-2630), 16 physical CPU cores (32 with hyper-threading), 32GB of RAM, dual NVIDIA Corporation GM200 GeForce GTX TITAN X GPUs, a 256GB SSD and Ubuntu Linux. Moreover, we used the ASU-Mayo Clinic polyp database[3] which currently is the largest publicly available dataset consisting of 20 videos with a total of $18,781$ frames and different resolutions up to full HD [31]. In these experiments, we implemented the system in Java, C++ and CUDA (for GPUs). We did not include any other data apart from the visual information, such as sensor data, etc., but this will be an important step for the future. For example, using results from a fecal blood test or temperature data will most probably increase the classification performance.

1) Detection Accuracy. To evaluate detection accuracy, we used the common standard metrics precision, recall and F1 score. We conducted a leave-one-out cross-validation to evaluate the system which is a method that assesses the generalization of a predictive model.

The system that we have developed allows us to use several different global image features for the classification. The more image features we use, the more computationally expensive the classification becomes. Also, not all image features are equally important or provide equally good results for our purpose. As a first step, we therefore need to find out which image features we want to use for classification. In order to understand which image features provide the best results, we generated indexes containing all possible features provided by LIRE [18]. These indexes were used for several different measurements and also for the leave-one-out cross-validation. Using our detection system, the built-in metrics functionality can provide information on the performance of different image features for benchmarking. Further, it provides us with separate information for every single image feature, as well as the late fusion of all the selected image features.

For our first test, we ran the detection with all possible image features selected. We then combined the reported values for true-positives, true-negatives, false-positives and false-negatives for all the runs, and calculated the metrics for the combined values. The single image feature that generally achieves the best score is CEDD, which is discussed in detail in [8]. Further, also the image features JCD, Edge Histogram, Rotation Invariant Local Binary Patterns, Tamura and Joint Histogram achieve very good values. The late fusion of all the image features even achieves slightly better results. However, it is impractical to do a late fusion of all these image features as the calculation, indexing and searching of all image features is computationally expensive. Therefore, we want to find a small subset of two image features, which provides optimal results despite minimizing the computational effort.

Based on the evaluation of different combinations of image features the image features JCD and Tamura seemed to be the best ones for our performance measurements. To assess the actual performance of the classifier combining these two image features, we ran the leave-one-out cross-validation over all available video sequences. With these settings, we achieve an average precision of 0.889, an average recall of 0.964 and an average F1 score value of 0.916. The problem with this average calculation is that different video sequences contribute values based on different numbers of video frames. If we weight the values contributed by every single video sequence with the number of frames in the sequence, we achieved an average precision of 0.9388, an average recall of 0.9850, and an average F1 score value of 0.9613. In other words, these results mean that we can detect polyps with a precision of almost 94%, and we detect almost 99% of all frames containing polyps. The detailed results compared to state-of-the-art systems are presented in Table 1. Furthermore, for the localization of the polyps in the frames, we reached an average precision of 0.3207, a recall of 0.3183 and a F1 score of 0.3195. These values are low in absolute terms and show how complex and difficult it is to make a multimedia system that is really useful for the medical doctors.

Obviously, more research is needed such as neural networks, more data, different classifiers, include humans in the loop, and methods have to be developed that can help to measure if performance is sufficient compared to the user needs. However, the multimedia community has to be aware that we cannot just apply our methods that we are used to use in this new field. Stated plainly, detecting cars or cats is not the same as detecting polyps or bleedings. For example, neural networks are conceptually easy to understand and lately large amount of academic research has been done on them. Results recently reported on for example the ImageNet dataset look quite promising [11]. Nevertheless, they have some negative aspects that make them less useful for the medical field [10]. First, training is very complicated and takes a long time. Our system has to be fast and understandable since we deal with patient data, and the outcome can differentiate between life and death. Therefore, a *black box* approach, that has difficulties to explain certain decision made, seems to be the second best way to solve a problem that has to be understood very well by all users. This can lead to serious problems in the medical field since it is not possible to evaluate them properly, and there will always be a chance that they completely fail without being aware of it [26]. The best way is still to understand the problem and then solve it. This of course comes with a challenge for the multimedia community. We have to test our current methods and most probably develop new, handcrafted algorithms and tools from scratch for this new field. A further problem of neural networks is that they require a lot of training data. In the medical field, this is a very important issue since it is hard to get data due to the lack of experts time (doctors have a very high workload) and legal and ethical issues for being able to share data among countries or even hospitals in the same country. Some common conditions, like colon polyps, may reach the required amount of training data for a neural network while other endoscopic findings, like for example tattoos from previous endoscopic procedures (black colored parts of the mucosa), are not that well documented, but still important to detect [28]. Finally, neural networks are not easy to design for probabilistic results. In a multi class decision based system, that is built to support medical doctors in decision making, the probability is an important information. Approaches with a better understanding of the problem will

[3] http://polyp.grand-challenge.org/site/Polyp/AsuMayo/

give a much more accurate probabilistic score that can be directly translated to the real world scenario [30].

2) Real-Time System Performance. One further requirement for the system and the medical field in general is scalability and execution performance. This requirement comes with some challenges like for example lack of actual hardware (it is in general hard to replace hardware or operating systems in hospitals due to security and system restrictions), not being able to use distributed systems and lack of funding for new hardware (e.g., Norwegian hospitals in 2016 still use Windows XP and Internet Explorer 6 even though funding is good). These restrictions makes it very challenging for researchers to develop efficient algorithms that are also scale able on the large amount of data that they will have to process. Therefore sophisticated methods are needed that run efficient in terms of speed and hardware need but at the same time achieve good performance. Based on our example system we present a experiment that shows how this challenges can be solved using multimedia systems knowledge and methods. For the experiments, we decided to use the configuration of the system that performed best in the accuracy experiment. In our use case of supporting doctors during live colonoscopies, it is important to reach real-time performance in terms of processing a video and several other input signal at the same time and reach a frame rate of not less than 30 FPS (output rate of current endoscopes). The performance of the *detection* is important, since the system should provide a result as fast as possible and not slower than 30 FPS making it usable for live applications. Figure 5(a) shows the detection sub-system performance in terms of FPS for the highest video resolution of 1920×1080. It depicts performance for all different detection algorithm implementations (Java, C++ and GPU) and different combinations of utilized hardware resources (from 1 to 32 CPU cores and none, 1 or 2 GPUs). For the full HD videos, the required frame rate of 30 FPS is reached using 8, 5 and 1 CPU cores in parallel for the Java, the C++ and the GPU implementations, respectively. Increasing the number of used CPU cores also increases the performance for all implementations, and the system reaches the maximum performance of 330 FPS with 2 GPUs and 25 CPU cores. A slight decrease of the performance can be observed for a high number of used CPU cores. This is caused by an increased overhead for context switching and competition for resource. Figures 5(b) and 5(c) show the detection sub-system performance in terms of FPS for the videos with smaller resolution. The maximum performance of 430 (for 856×480 resolution) and 453 (for 712×480 resolution) FPS is reached using 2 GPUs and 18 and 16 CPU cores. For localization which is more computationally expensive (plots not shown), the maximum performances observed are 129, 246 and 283 FPS for 1920×1080, 856×480 and 712×480 resolutions, respectively.

The outcome of these experiments clearly shows that our system can reach real-time requirements for the video processing and still has processing power left which can be used to process other input data at the same time, for example, sensor or patient records data, etc. A number of complex features can be added into the detection and the localization sub-systems. This will increase the system's detection and localization accuracy, and at the same time, keep its ability to perform in real-time. Moreover, it can also be used to process several data streams simultaneously in real-time and significantly reduce the examination time of the VCE videos for the medical experts. The time reduction lies around 5-10 times depending on type of input data like for example video resolution, frame rate and sensors used. Our evaluation also shows, that this is a very complex topic and requires methods and technologies from several different multimedia research directions, e.g., signal processing, multimedia systems, information retrieval, etc.

6. OUTLOOK AND CHALLENGES

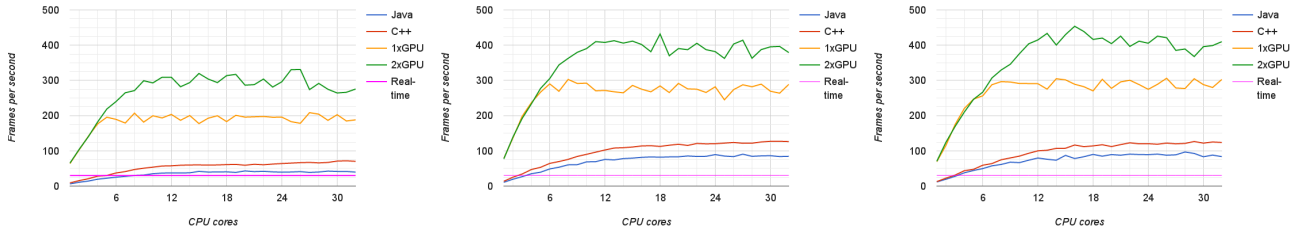
With 2.8 million cancer cases diagnosed in the GI system per year with a mortality rate of about 65%, we have the best motivation to perform research in the proposed area. The GI example that we used in this paper is only the tip of the iceberg of unsolved problems in the health care sector. By exposing more unexplored multimedia research questions, researchers can reveal a huge potential to save lives by combining the medical and multimedia research areas. Our aim is to raise awareness that (i) multimedia research can do a lot for and learn a lot from the field of minimally invasive medicine, (ii) interdisciplinary research in this field leads to immediate benefits, and (iii) we have only scratched the surface with our efforts.

In our experience, medical experts are open to new multimedia applications in their fields. We experienced that doctors are willing to spend a lot of time and effort into supporting such research, as it ultimately has the potential to make their daily routine more efficient, and they will have more time to focus on the patients themselves. Especially, since we live in a time where handling multimedia is part of everyone's lives, medical experts wonder why the same functionality that they can use in YouTube, Flickr and Twitter cannot be applied to their own medical field. The main reasons that we identified are that first of all the computer vision and medical imaging community that work mainly on this problems is not interested in the *whole multimedia life cycle from start to end*, i.e., from the content creation, analysis to content usage by the actual users. Second and most important, it is a problem within our own community. It is much more convenient to download pictures from Flickr or videos from YouTube and categorize and use them in research, especially as many can identify themselves as social media users. However, working with medical data involves organizational challenges like *seeking and maintaining contact with medical experts*, understanding their problems, as well as getting used to often unpleasant or even content that causes a disgust response until a researcher is habituated in working in the area. Nevertheless, if we – the multimedia community as a whole – would be more brave to tackle these problems, we could actually help to save lives, make patient examinations less uncomfortable and help to save money and time spent in the health care system for daily routines instead of research. These are possibilities for societal impact that surely are appealing for both, researchers as well as global citizens. Last but not least, being able to look back seeing that our multimedia research helped to save lives is bearing more weight than being able to say we can classify cats, cars or beautiful holiday pictures.

6.1 Open Challenges

Our EIR system has preliminarily shown how multimedia tools can impact greatly health care systems. Nevertheless, there are still many open challenges that need to be faced through a multidisciplinary approach where multimedia methods will have to play a key role. Challenges include but are not limited to:

1) Exploiting domain expert knowledge to improve automated methods performance. Most of the methods (including the ones described in this paper) devised for supporting medical investigations in analysing visual data content are still predominantly based on learning distributions of low-level and middle-level (recently using deep learning approaches) visual features. While this has proved to achieve good performance in many computer vision applications, there are cases, especially in the medical domain, where relying on visual appearance might fail since processing visual data content requires specific expertise. This is the case of endoscopy videos where the reliability of the outcome mainly depends on the examiner's expertise. Our hypothesis is that, for a real break-



(a) Videos with a resolution of 1920×1080. (b) Videos with a resolution of 856×480. (c) Videos with a resolution of 712×480.

Figure 5: The performance of the detection sub-system in terms of FPS varying the number of CPU cores, the resolution of the videos and the detection algorithm. The maximum performances observed are 330, 430 and 453 FPS for 1920×1080, 856×480 and 712×480 resolutions.

through in medical image analysis, automated methods need to exploit jointly perceptive elements (visual features) and semantic factors (domain knowledge). This explains why in the medical domain relying only on image processing and computer vision methods will lead to a dead end. Instead, a multidisciplinary approach operating on multimodal data is necessary. Nevertheless, exploiting high level knowledge in computer vision methods poses several challenges from how to extract and model effectively domain expert knowledge to how to include such semantics into machine learning methods.

2) Automated report systems. A significant part of a medical professional's time is spent for preparing reports after procedures and examinations. Multimedia research can significantly support this phase by collecting all patient and examination data and by providing automatically summaries able to convey key information of the performed procedures including media fragments, e.g., video frames with detected objects, audio speeches describing colon visual features, etc. Such distilled media needs also to be interlinked with detailed information on treatments, medication for a holistic view of patients. These report will also be extremely useful for training medical experts: through multimedia enriched reports, medical doctors in training can learn based on real data according to case-based teaching and problem-based learning strategies. The multimedia field has tackled over the years, the problem of multimedia summarization for automated report generation, but such research is still at its infancy since methods developed so far are able to process only one type of media at a time (hence do not take full advantages from the richness of multimodal data). However, the most important limitation of multimedia research in this direction is the lack of generalization capabilities; in fact, most approaches cannot be applied to domains different from the ones they were devised for. To overcome these limitations, one solution we believe is worthwhile to investigate is to build automated multimedia summarization methods with a semantic nature exploiting domain ontologies, which can play an important role in the medical multimedia analysis where the data complexity and heterogeneity make the task very challenging.

3) Integration and fusion of unstructured and heterogeneous data. Beside visual data, other (equally important) information (e.g., blood pressure, temperature, breathing, oxygen levels) are recorded during examinations, which, if suitably fused to visual data content may significantly enhance procedures' outcome. An additional, and semantically rich, data source that can be exploited is recordings of medical experts spoken comments during examinations. Indeed, surgeons often describe verbally the procedure by giving details on what they see to other doctors and to issue commands and requests to the medical team. Although audio generated during procedures is a valuable source of information to train both automated methods and young doctors, it is rather unstructured and noisy and, as such, it demands for specific text mining methods approaches to distill the key information and to map it to a structured data form. Under this scenario, the semantic web may be a powerful tool for integration of such heterogeneous multimedia data. Once, heterogeneous data are all modeled using a shared formalism, visualization approaches are envisaged to present fused information in order to support medical staff, by enhancing the examination experience, for diagnosis.

4) Patient context information. Typically, health issues affect patients beyond their immediate treatment, and there are very often preceding correlated events before treatment is necessary or a health related issue is diagnosed. Therefore, health issues do not appear suddenly or as isolated events, but come in a rich context, which is largely exploited by medical doctors for diagnosis and treatment. Such context includes patients' mobility, eating habits and changes, etc. To this end, multimedia research can play an important part in developing smart wearable body sensors (and algorithms to analyze their data) that can collect routinely all such information and share with medical staff.

5) Building a knowledge base. A large collection of multimedia including videos, audio streams, sensor readings and patient records will represent a priceless knowledge base for approaches like case based reasoning and/or large empirical studies on treatments. Nevertheless, sharing such knowledge base opens up issues in privacy and data security, that, if successfully addressed, will enable the increase of such knowledge base (since many medical people will share their data), thus leading to large scale benefits in health care. To effectively address protection and reliability issues, multimedia researchers should investigate secure communications and processing through a deep interaction between signal processing, networking, and cryptography.

6) Interlinking information from different modalities. Besides endoscopic and minimally invasive surgery, there are other diagnosis systems like X-Ray, ultrasonic or MRT data from patients. Surgeons would greatly benefit from synchronized spatial information on multiple modalities to be able to investigate abnormalities from different angles. Now, all interlinking of diagnostic data from multiple modalities has to be done manually. This shows that there exists a huge need for algorithms and applications that can combine these different types of media automatically and efficient. For example, the information collected from a standard colonoscopy with a video from a capsular colonoscopy and CT colonography (virtual colonoscopy that uses special X-ray equipment) could lead to a higher detection rates and better patient survival probabilities.

7) Simplifying handling of multimedia. With today's tools, everyone is used to access multimedia everywhere and manipulate and share multimedia data with the tip of a finger. In the medical domain, software systems have a comparably long life span, and it has to be thoroughly tested before they can be applied in a hos-

pital setting. Therefore, we need sustainable interactive tools and ways of interactivity that do not wear off as fast as they did in the last decade. Multimedia researchers have the knowledge and are needed to help creating such systems that fulfill the user needs but also to develop the algorithms that are the basis of such systems such as content retrieval, etc. This is especially important since most of the standard algorithms for object or concept detection will most probably not work in the medical field, which we experienced in the begin of our research when we tested a lot of state-of-the-art methods like for example histogram of oriented gradients or structured output tracking with kernels, etc. We believe that this is mainly caused by differences in the multimedia data provided (videos and images show completely different content, quality of the data, needs of the users, etc.).

8) Test data sets and challenges. There are already workshops, challenges and whole conferences dedicated to the topics of medical information and multimedia systems. However, just like in the multimedia community, we have to move forward to build and maintain an over-critical mass of test data including ground truth and annotations, and usage scenarios that are recent enough, i.e., recorded with up-to-date sensors and annotated thoroughly based on current medical standards and state-of-the-art. This is not only a research, but also a legal and societal, challenge as medical data is always personal and especially if it includes a patient context or long term records it is hard to anonymize. This requires not only sophisticated annotation systems, but also algorithms for unsupervised and semi-supervised learning. Furthermore, algorithms that can help to anonymize or watermark content to protect data are needed. Apart from the algorithms to analyze the data this part also needs motivated and dedicated people that contact hospital key personnel and doctors, and play a pioneering role in establishing a good data basis by collecting, annotate and make data public available.

9) Acting in concert. The greatest challenge of all, however, is to act in concert, as an interdisciplinary community. Medical experts bring in the data as well as the domain knowledge. Legal experts find ways how to deal with privacy and data security aspects from a legal and societal point of view. Companies supplying medical equipment must open up for collaboration and research beyond their own research departments. Last but not least, the multimedia community must bring in its knowledge as a core discipline, but also as a research field which historically involved other disciplines like computer vision, machine learning, interactive systems, networking, data warehousing, speech recognition, information retrieval, data mining and software engineering. The biggest task that the multimedia community faces is most probably to break the ice. Medical experts often do not know what is even possible with the data they have. Therefore, the responsibility lies in the hands of the multimedia researchers to build bridges. For example, we went to hospitals and asked for meetings with doctors to show them what we can do. Once they saw the possibilities, they were willing and very motivated to contribute with knowledge, data and new ideas. To address all these challenges, an interdisciplinary team is necessary as the problems goes far beyond visual analysis, information retrieval and annotation. It is also a multimedia area where it is essential to involve researchers from different areas like interactive system, multimedia systems and speech recognition in a specialized domain, ontologies, data mining and machine learning, sensor fusion, and synchronization of data from different modalities.

6.1.1 Possible Research Projects

We encourage the multimedia community to be open minded and help to tackle the challenges in this new field. It is important to be aware that we cannot just keep on annotating social videos, and then expect that medical technology companies can transfer these technologies to the medical use case. Therefore we need specific approaches for the field of medical multimedia.

In the sense of getting more into detail, we want to point out the more immediate and concrete challenges in this field by proposing three different research project topics and relevant research questions making for multiple challenging and interesting PhDs.

1) How can we identify and track abnormalities in a live endoscopic video? While our prototype did experiments on doing exactly that, there are fields beyond polyps as well as an opportunity to reduce manual input. Going beyond polyps would mean to identify cancerous tissue, inner injuries, bleeding, scars, fractures, and so on. This goes well with finding the current position and rotation of the camera within a patients body, i.e., by sensor fusion and asks for new and multimodal tracking algorithms taking camera movement into account. Medicine needs very high recall, but false alarms can be very costly not to mention extremely upsetting for the patients. Multimedia that detects concepts or events in YouTube videos is just not held to these kinds of standards.

2) How can we pre-prepare the final report on the surgery? As reporting takes a lot of a surgeons time, any step in this direction would be immediately beneficial for medical experts and patients alike. This actually involves several multimedia disciplines. Many surgeons direct and inform their team during a surgery by short, spoken announcements like *"Here, we've got the first polyp."*, *"Electro scalpel!"* or *"This one looks particularly odd."*. With speech recognition and synchronization with a video stream, the video can be segmented, relevant parts can be found and media for a final report can be suggested in addition with recommending relevant text passages from earlier reports of similar cases. The systems need to be able to optimize not for correct predictions, but for what humans need to know in order to make decisions. One approach is to fuse many slightly different algorithms so that the typical mistakes of one algorithm do not accidentally dominate.

3) How can we share, annotate and educate? While of course many would like to see a YouTube or Flickr like social media network for medical experts, it is simple not possible as the number of experts is limited and not everyone can be expected to be an active contributor to such a network. However, especially senior surgeons are skilled in creating videos, books or training materials and communicating them to trainees or colleagues to exchange knowledge. Still they lack tools for that. Critical for such a venture would be interdisciplinary work in (i) interactive multimedia like annotation, share, and interlinking of content, (ii) security and encryption for making sure the data stays safe, (iii) knowledge based systems as ontologies and structured knowledge plays a huge part in that, and (iv) multimedia systems, as all the data has to be handled, transferred, streamed, encoded etc.

6.1.2 First Steps

While we stressed the fact that working with medical data and medical experts is crucial for moving forward with research in the medical domain, we also acknowledge that interdisciplinary work is hard to start. What we found most important in our project is to build a working relationship with medical doctors who are personally interested in *making things better*. The VIPs for such interdisciplinary projects are senior surgeons, who are actively training new surgeons, as they (i) have experience in sharing knowledge, (ii) have access to a lot of data, (iii) are extremely good in specifying problems and very competent in working out solutions, and (iv) have influence in terms of the hospital organization.

In our experience, it takes some time for PhD students to build

awareness of the field to a level, where we could work efficiently on the problem. At the begin, we organized that the PhD students attended live surgeries, watched and discussed surgery videos and reports with senior surgeons as well as trainees, and participated in regular meetings for questions and answers that were raised in this learning period. Within this starting period, in parallel with building up the knowledge, it is in general a good idea to expand the data available throughout the research project. Besides building on public data sets like the ASU-Mayo Clinic polyp database [31], we suggest to work out a scheme to obtain recent multimedia data from the before mentioned necessary contacts. This typically involves legal and organizational issues including but not limited to (i) a mutually agreed upon anonymization routine for the data, (ii) a non disclosure agreement of the participating organizations and involved people, as well as (iii) a specialized setup to make sure the data stays safe and protected during transport and in storage at the research institution.

7. REFERENCES

[1] European Commission forms EUR2.5bn big data partnership. http://www.pmlive.com/blogs/digital_intelligence/archive/2014/november/european_commission_forms_2.5bn_big_data_partnership.

[2] Obama's big data plans: Lots of cash and lots of open data. http://gigaom.com/cloud/obamas-big-data-plans-lots-of-cash-and-lots-of-open-data/.

[3] 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 *Proc. of MMSYS*, 2015.

[4] L. A. Alexandre, J. Casteleiro, and N. Nobreinst. Polyp detection in endoscopic video using svms. In *Proc. of PKDD*. Springer, 2007.

[5] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino. Texture-based polyp detection in colonoscopy. In *BFM*. 2009.

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

[7] 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.

[8] 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*, 2008.

[9] 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 *MDAIS*. Springer, 2008.

[10] C. Chin and D. E. Brown. Learning in science: A comparison of deep and surface approaches. *Research in science teaching*, 37(2), 2000.

[11] 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*. IEEE, 2009.

[12] M. M. Francisco, B. S. Terry, J. A. Schoen, and M. E. Rentschler. Intestinal manometry force sensor for robotic capsule endoscopy: An acute, multipatient in vivo animal and human study. *Trans. on Biomedical Engineering*, 63(5), 2015.

[13] P. Halvorsen, S. Sægrov, A. Mortensen, D. K. Kristensen, A. Eichhorn, M. Stenhaug, S. Dahl, H. K. Stensland, V. R. Gaddam, C. Griwodz, and D. Johansen. Bagadus: An integrated system for arena sports analytics – a soccer case study. In *Proc. of MMSys*, 2013.

[14] 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*, 2007.

[15] 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. *NE Journal of Medicine*, 362(19), 2010.

[16] J. Kang and R. Doraiswami. Real-time image processing system for endoscopic applications. In *Proc. of CCECE*, 2003.

[17] B. Li and M.-H. Meng. Tumor recognition in wireless capsule endoscopy images using textural features and svm-based feature selection. *ITBM*, 16(3), 2012.

[18] M. Lux. LIRE: open source image retrieval in java. In *Proc. of MM*. ACM, 2013.

[19] A. Mamonov, I. Figueiredo, P. Figueiredo, and Y.-H. Tsai. Automated polyp detection in colon capsule endoscopy. *MI*, 33(7), 2014.

[20] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_-Innovation/Big_data_The_next_frontier_for_innovation.

[21] McKinsey Global Institute. The big-data revolution in US health care: Accelerating value and innovation. http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care.

[22] B. Münzer, K. Schoeffmann, and L. Böszörmenyi. Detection of circular content area in endoscopic videos. In *Proc. of CBMS*, 2013.

[23] B. Münzer, K. Schoeffmann, and L. Böszörmenyi. Improving encoding efficiency of endoscopic videos by using circle detection based border overlays. In *Proc. of ICME*, 2013.

[24] B. Münzer, K. Schoeffmann, and L. Böszörmenyi. Relevance segmentation of laparoscopic videos. In *Proc. of ISM*, 2013.

[25] R. Nawarathna, J. Oh, J. Muthukudage, W. Tavanapong, J. Wong, P. C. De Groen, and S. J. Tang. Abnormal image detection in endoscopy videos using a filter bank and local binary patterns. *Neurocomputing*, 144, 2014.

[26] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv preprint arXiv:1412.1897*, 2014.

[27] 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*, 2014.

[28] J. Schmidhuber. Deep learning in neural networks: An overview. *NN*, 61, 2015.

[29] 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. *BJOC*, 106(5), 2012.

[30] D. F. Specht. Probabilistic neural networks. *NN*, 3(1), 1990.

[31] N. Tajbakhsh, S. Gurudu, and J. Liang. Automated polyp detection in colonoscopy videos using shape and context information. *Trans. on MI*, 35(2), 2015.

[32] The New York Times. The $2.7 Trillion Medical Bill, 01, Jun, 2013.

[33] The New York Times. The Weird World of Colonoscopy Costs, 06, Sept, 2013.

[34] The Telegraph. 'spider pill' offers new way to scan for diseases including colon cancer, 11, Oct, 2009.

[35] 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. *JOG*, 101(2), 2006.

[36] 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.

[37] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Computer-aided detection of retroflexion in colonoscopy. In *Proc. of CBMS*, 2011.

[38] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Near real-time retroflexion detection in colonoscopy. *JBHI*, 17(1), 2013.

[39] 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. *JBHI*, 18(4), 2014.

[40] 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), 2015.

[41] 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, 2012.

[42] 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*, 2014.

Paper VIII

A Holistic Multimedia System for Gastrointestinal Tract Disease Detection

A Holistic Multimedia System for Gastrointestinal Tract Disease Detection

Konstantin Pogorelov
Simula Research Laboratory, Norway
University of Oslo, Norway

Sigrun Losada Eskeland
Bærum Hospital, Norway

Thomas de Lange
Bærum Hospital, Norway
Cancer Registry of Norway

Carsten Griwodz
Simula Research Laboratory, Norway
University of Oslo, Norway

Kristin Ranheim Randel
Cancer Registry of Norway
University of Oslo, Norway

Håkon Kvale Stensland
Simula Research Laboratory, Norway
University of Oslo, Norway

Duc-Tien Dang-Nguyen
Dublin City University, Ireland

Concetto Spampinato
University of Catania, Italy

Dag Johansen
UiT–The Arctic University of Norway

Michael Riegler
Simula Research Laboratory, Norway
University of Oslo, Norway

Pål Halvorsen
Simula Research Laboratory, Norway
University of Oslo, Norway

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.

CCS CONCEPTS

• **Information systems → Multimedia information systems**;

KEYWORDS

Interactive; Medicine; Gastrointestinal Tract; Medical Multimedia System; Performance; Evaluation

This work is founded by the Norwegian FRINATEK project "EONS" (#231687).
Contact author's address: Konstantin Pogorelov, Simula Research Laboratory, Oslo, Norway, email: konstantin@simula.no .

MMSys '17, June 20–23, 2017, Taipei, Taiwan

ACM ISBN 978-1-4503-5002-0/17/06.
https://doi.org/http://dx.doi.org/10.1145/3083187.3083189

ACM Reference format:
Konstantin Pogorelov, Sigrun Losada Eskeland, Thomas de Lange, Carsten Griwodz, Kristin Ranheim Randel, Håkon Kvale Stensland, Duc-Tien Dang-Nguyen, Concetto Spampinato, Dag Johansen, Michael Riegler, and Pål Halvorsen. 2017. A Holistic Multimedia System for Gastrointestinal Tract Disease Detection. In *Proceedings of MMSys '17, Taipei, Taiwan, June 20–23, 2017,* 12 pages.
https://doi.org/http://dx.doi.org/10.1145/3083187.3083189

1 INTRODUCTION

The human gastrointestinal (GI) tract can potentially be affected by various abnormalities and diseases, including colorectal cancer (CRC) which is a major health issue world wide. For the case of CRC, an early detection is crucial for survival, and several studies demonstrate that a population-wide screening program improves the prognosis and even reduce the incidence of CRC [23]. As a consequence, in the current European Union guidelines, screening for CRC is recommended for the population over 50 years of age [57].

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 as a work-up tool after other positive screening tests [33]. However, endoscopies are invasive procedures and may be of great discomfort for patients. Long-lasting training of physicians or nurses is required to perform the examinations. They are performed in real-time and are challenging to scale to a larger population. Additionally, the procedure is expensive. In the US, for example, the colonoscopy is the most expensive cancer screening process with annual costs of 10 billion dollars ($1100/person) [55], and with a time consumption of about one medical-doctor-hour and two nurse-hours, per examination.

In this respect, we propose a scalable, real-time disease-detection system for the GI tract. The idea is to assist endoscopists (physicians highly trained in the procedure) during live examinations. Additionally, alternatives to traditional endoscopy examinations have recently emerged with the development of non-invasive endoscopy capsules (WVCs). The idea is a pill-sized camera (available from vendors such as Given and Olympus), that is swallowed and

then records a video of the entire GI tract. The challenge in this context today is that medical experts still need to view the video in a non-scalable way. Our system should provide a scalable system that can be used as a first-order population screening system where the WVC-recorded video is used to determine whether a traditional endoscopic examination is needed or not.

The system presented in this paper is designed to support detection of a wide range of diseases, but our initial focus is on colorectal polyps and a small subset of other diseases. Polyps are specifically relevant because they are known precursors of CRC (see for example figure 2 and 3). The reason for starting with this scenario is that most colon cancers arise from benign, adenomatous polyps containing dysplastic cells that may progress to cancer. Detection and removal of such polyps prevents the development of cancer. Thus, the risk of getting CRC the following 60 months after a colonoscopy depends largely on the endoscopists ability to detect polyps [25].

In the context of object or pattern detection and tracking in images and videos, there has been a lot of research, and current systems are good at detecting human faces, cars, logos, etc. However, detecting diseases in the GI tract is very different from detecting objects like logos or cars. The GI tract can potentially have a wide range of lesions visible on endoscopy, as well as findings associated with benign/normal or man-made lesions. This leads to necessity of distinguishing between multiple classes of diseases, including findings with high level of visual similarity. In this scenario, both high precision and recall are of crucial importance, but also is the often ignored system performance in order to provide live feedback because medical personal is assisted most efficiently while they perform the examination. The most recent and most complete related work is the polyp detection system Polyp-Alert [61], which can provide near real-time feedback during colonoscopies. However, it is limited to polyp detection, and it is not fast enough in the case of live examinations.

To further aid and scale such examinations, we have developed EIR [47], an efficient and scalable information retrieval system for medical data like videos and images. The system supports endoscopists in the detection and interpretation of diseases in the GI tract. In this paper, we provide more detailed description of our EIR system, we greatly extend the evaluation, and we also introduce localization. The main objective of the system is to develop both (i) a live-system assisting the visual detection of diseases during colonoscopies, and (ii) a future fully automated first line screening for CRC using WVCs. Both goals pose strict requirements for the accuracy of the detection in order to avoid false negative findings (overlooking a disease) as well as low resource consumption. The live assisted system also introduces a real-time processing requirement (defined as being able to at least process 25 frames or images per second). In this paper, the initial prototype of our system is presented. This is built by combining filters using machine learning, image recognition and extraction and comparison of global and local image features. The system will be extended to support detection of multiple abnormalities and diseases of the GI tract by training the classifiers using different datasets. We evaluate our prototype by training classifiers that are based on the different image recognition approaches. It is important to point out that these classifiers can also process other input like for example sensor data. We also test the generated classifiers with different diseases and thereby evaluate the different approaches for feasibility of colonic polyp recognition and localisation.

The initial results from our experimental evaluation show that: (i) the detection and localisation accuracy can reach the same performance or outperform other current state of the art methods, (ii) the system performance reaches real-time in terms of video processing up to high definition resolutions.

The rest of the paper is organized as follows: we present related work in section 2. This is followed by a description of the complete system in section 3. After that, we present a detailed evaluation of the whole system in section 4, and we discuss in section 5 two cases where our system will be used in two medical examinations. Finally, we draw the conclusion in section 6.

2 RELATED WORK

To the best of our knowledge, no related work that presents a complete multimedia system for analysing the whole GI tract in real-time exists. The complete system covers the entire pipeline from data capture to live detection feedback, and has to fulfill many requirements. These requirements include (i) high detection accuracy, (ii) real-time processing to support live examinations like colonoscopies, (iii) efficient resource utilization to allow massive scale using WVCs, and (iv) expandability to allow the system to support new diseases.

Detection of diseases in the GI tract has mostly focused on polyps. This is most probably due to the lack of data in the medical field and polyps being a condition with at least some data available [30]. Automatic analysis of polyps in colonoscopies has been in the focus of researchers for a long time and several studies have been published [58, 59, 62]. However, not many systems are able to do real-time detection or support doctors by computer aided diagnosis during colonoscopies in real-time. Furthermore, all of them are limited to a very specific use case, which in the most cases is polyp detection for a specific type of camera.

Several algorithms, methods and partial systems have been proposed and have achieved results in their respective testing environment that are promising. However, most of the research conducted in this field uses a rather small amount of training and testing data, making it difficult to generalize the methods beyond the specific dataset and test scenarios. In the [47] paper, we presented a summary of the detection performance and speed properties of the most relevant approaches in colonoscopy and polyp detection. The conducted search through the relevant publications [3, 4, 9, 24, 26, 28, 34, 60, 61, 63] showed that different researchers provide different metrics for measuring the performance and use different datasets for training and testing. Moreover, almost all of the researches focus on polyps only.

The Polyp-Alert approach from Wang et al. [61] is the most recent, most complete and best working in the field of polyp detection. It is able to give near real-time feedback during colonoscopies. The system can process 10 frames per second and uses visual features and a rule-based classifier to detect the edges of polyps. Further, they distinguish between clear frames and polyp frames in their detection. The researchers report a performance of 97.7% correctly detected polyps, based on their dataset which consists of 52 videos

taken from different colonoscopes. Unfortunately, the dataset is not publicly available and therefore a detection performance comparison is not possible.

Mamonov et al. [34] presented an algorithm for a binary classifier with pre-selection to detect polyps in the colon. The used assumption is that polyps can be generalized as protrusions (something that bumps out) that are mostly round in shape. The researchers report a sensitivity of 81.25% per polyp at a specificity of 90%. The sensitivity of the algorithm with regards to single input frames is significantly lower and only reaches 47%. The length of an input sequence varied between 2 and 32 frames and a total of 16 sequences were tested. The false positive rate on the total of 18,738 frames not containing a polyp was 9.8%. Assuming that it is usual to have multiple frames available for a single polyp, these numbers seem quite promising. With this method, the time a specialist has to spend on evaluating video data could be reduced by about 90%.

A similar approach is presented by Hwang et al. [24]. This approach also focuses on shape, in particular on ellipses, which is a common shape for a polyp. Using this method, a frame is first segmented into elliptical regions by a watershed-based image segmentation algorithm. These regions and corresponding ellipse edges are then evaluated for matching of curve direction, curvature, edge distance and intensity. 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. To evaluate the method, a video sequence with a frame rate of 15 fps has been processed. Out of 27 available polyp shots (frames containing a polyp), 26 were detected correctly with a total of 5 false-positives. Similar to [34], the authors assume that multiple frames are available for one polyp and that a certain number of false-negatives is acceptable in order to balance the number of false-positives. The correctness of this assumption depends strongly on the frame rate of the camera that is used for recording the video.

Another recent approach related to our approach and not limited to polyps is presented by Nawarathna et al. [39]. In the paper, the authors describe a method to detect abnormalities like bleeding, but also polyps in colonoscopy videos. The authors use a texton histogram of an image block. The authors report a 91% recall and a 90.8% specificity for colonoscopy images.

Other papers that discuss how to improve performance of endoscopic surgeries in general (not colonoscopy) are for example [36–38]. In these papers, the authors report their method for detecting the circular content area that is typical in endoscopic videos. Furthermore, they present their method for relevance segmentation in endoscopic videos. The methods seem to be very useful in terms of archiving and saving storage space.

Since neural networks (NNs) are commonly used nowadays, they are also discussed for automatic analysis of GI tract videos. NNs are conceptually easy to understand and lately large amounts of academic research has been done on them. Results recently reported on, for example, the ImageNet dataset look quite promising [13]. Nevertheless, they have some negative aspects that make them less useful for our use case [10]. First, NNs are a *blackbox* approach. This can lead to serious problems in the medical field since it is not possible to evaluate them properly, and there will always be a chance that they completely fail without being aware of it [40]. Further, training of NNs is complicated, takes a long time and requires a lot of training data. In the medical field, this can be a challenge since it is hard to get data due to the lack of experts' time and because of legal and ethical issues. Some common conditions such as colon polyps may reach the required amount of training data for a NNs while other findings, like tattoos from previous endoscopic procedures, are not that well documented but still interesting to detect [48]. Finally, NNs are not easy to design for probabilistic results. In a multi-class decision-based system that is built to support medical doctors in decision making, the probability is an important information to help them finding a decision. Approaches with a better understanding of the problem give a much more accurate probabilistic score that can be directly translated to the real world scenario [50].

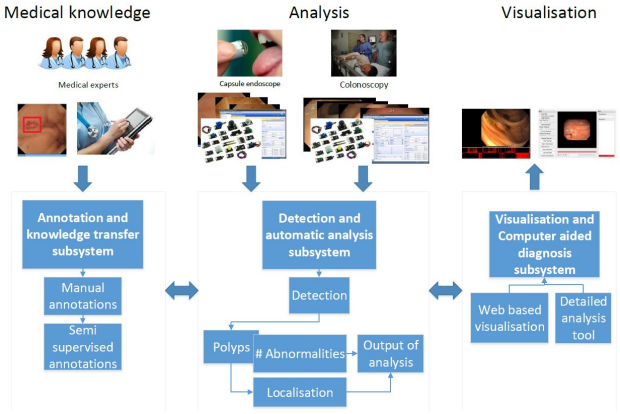


Figure 1: System overview with the three main subsystems: annotation, detection and automatic analysis and visualization.

In summary, a lot of related work with many interesting approaches for polyp detection exists. However, they (i) are either too narrow for a flexible, multi-disease detection system, (ii) have been tested on a too limited datasets not showing if the methods would work in a real scenario, or (iii) provide a too low performance for a real-time system or authors have ignored the system performance aspect in their evaluations altogether. To the best of our knowledge, our system is the first that aims at total flexibility in terms of diseases that can be detected, and at the same, time focuses on the performance and the evaluation of it.

3 BASIC IDEA OF THE SYSTEM

The objective of the system is to support doctors in GI tract disease detection, both as a live examination system and as an offline system for WVCs. Its main requirements are already listed in section 2, but it also has to be easy to use. Figure 1 gives an overview of the whole system. It consists of three main parts: the annotation subsystem, the detection and automatic analysis subsystem and the visualization and computer aided diagnosis subsystem.

3.1 Annotation Subsystem

It is well known that training data is very important for a classification system that relies on machine learning techniques. In the

medical field, both the time of the experts and available data are very limited. Even when experts' time can be acquired, the quality of annotations depends on their experience and concentration [17]. For each image or video, a patient consent has to be collected before research can be done, making it a very cumbersome task. The purpose of the annotation subsystem is therefore to efficiently collect training data for the detection and automatic analysis subsystem.

For example, in a single WVC procedure, there are several 100, 000 images per examination, and a very experienced endoscopist needs between one and several hours to view and analyze all the video data [29]. Due to this, it is important to develop automatic methods that can reduce the burden on physicians and speed up the process of video analysis. We therefore also developed an efficient semi-automatic annotation subsystem [2]. This tool makes it easy for doctors to annotate and provide data to the system. The manual annotations of the doctors are combined with semi-automatic methods that extend the provided data. Our semi-automatic process reduces the time that time physicians spend on annotating. Instead of annotating every frame, they can provide annotations on a single frame of an image series or video. They identify abnormalities, mark a region of interest and tag it accordingly. The automatic step [2] uses this information to track the regions of interest on subsequent as well as previous frames. Due to the fact that the medical doctor is usually located in a hospital with security restrictions, the implementation of the software is done with standard web technologies which do not require any installation at the hospitals systems. This also includes the storing of all information on the systems side and moves the responsibility of maintaining the system and data integrity from the user to the system. Besides getting data for the system to enable automatic screening, the annotation subsystem makes it possible to use the annotated videos in a medical video archive for surgical documentation or teaching purposes.

3.2 Automatic Detection Subsystem

The subsystem for detection and automatic analysis is designed in a modular manner, so that it can be easily extended to additional diseases, to new subcategories of a disease, as well as newly requested information, such as determining a polyp's size. At the moment, the subsystem consists of two parts, the detection subsystem that detects frames containing irregularities, and the localisation subsystem that localises the position of the irregularity within a detected frame.

3.2.1 Detection Subsystem. The detection subsystem detects whether a frame contains an irregularity, without any indication of a position of this irregularity in the frame. The detection of specific abnormality type can be performed after the initial training of the detection subsystem using previously collected training frames set. All frames that are used in training are divided into two disjoint sets. These two sets contain example images for abnormalities and images without any abnormality. Each of these sets can be seen as the model for a specific disease.

The detection subsystem supports a hierarchical concept of models and sub-models. This does allow it to, for example, first detect a polyp and then distinguish between a polyp posing a low or high risk of developing into CRC using the *NICE* classification [22]. To compare and determine the abnormalities in a given frame, we use global image features. In previous work [45], we showed that, in case of only detecting whether a frame contains an irregularity or not, global features can outperform local features, i.e., at least reach the same results with respect to detection and significantly outperform local features in terms of processing speed.

The whole system is built using the Lire [31] open source library for content-based image retrieval, written in *Java*. This library provides a comprehensive set of tested algorithms to extract a variety of global image features. It allows us to experiment with a wide range of global image features for detecting or clustering video frames from colonoscopy or WVC videos. Lire uses *Lucene indexes* [16] for storing and searching image feature data.

Indexing. The index structure is field- and row-based. Each row is defined by its fields, e.g., the image path, the binary values for the feature or the hash representation of the feature, etc. The number of fields and their size are variable depending on the number or type of feature. All feature values are stored as byte representation of the respective feature vector as well as a text field containing hash values from a random projection hashing [31] approach.

The hashing approach is based on locality sensitive hashing [31] (LSH). The main idea is to use multiple random hash functions to hash the values of the features giving the same hash values for the similar images. This is done by a linear projection in random directions of the hash functions in the feature space of the image. The created hash codes are ineffective and a large number of hash tables is needed to achieve a reasonable search quality, but compared to the increased speed of the algorithm these are minor disadvantages that can be ignored [49].

We use a hash function $h(v) \in \{0, 1\}$, which is defined for a histogram v as $h(v) = sgn(v \cdot r)$, where sgn is the sign function (extracts the sign of a real number) and r is a random vector with uniformly distributed elements r_i with $-w \leq r_i \leq w$. n hash functions are combined as a bit string in one single hash value $H(v) < 2^n$. For indexing m hash values, m functions $H_j(v)$, $0 \leq j < m$ are generated.

The parameters for the hashing-based approximate indexing are chosen based on evaluations on an image dataset consisting of 10^5 images. To achieve a good performance for precision and search time, the parameters have been set as following: $w = 2$, $n = 12$, and $m = 150$. This leads to a significant speed-up and at the same time, to a good trade-off between search time and precision.

Search. The search for an image that we use in our search-based algorithm is performed on the fly on the previously created indexes. For each image, a term-based query from the hashed feature values of the query image is created, 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 [54], which is computed by taking the ratio of the number of elements that intersect and the union of the elements:

$$f(A, B) : [0, 1]^n x [0, 1]^n \rightarrow \mathbb{N} = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$

A smaller distance between an image in the index and the query image means a better rank [54]. The final ranked list is used in the

classification step. To be able to classify an image efficiently, two important aspects have to be considered: the selected features and the feature combination.

Feature Selection. Different features have different properties, and they are therefore useful in different scenarios. To make the search-based classifier fast and accurate, we have to decide which features we want to use for a specific use case, because a random selection of global features and random combinations of feature can lead to negative results for the classification or search task. Badly chosen feature combinations can introduce noise (if too many features are combined and some of them do not add any information to the classification problem) or make the search slow (if the index is very big because of too many used global features). A lot of work has been performed in the field of feature selection, and different machine learning techniques were utilized for it [35]. For example, an information gain (IG) attribute evaluation, which computes the information gain of a given feature with respect to the classification problem to determine which feature gives the most information [12]. Another example is the SVM attribute evaluation, which ranks the variables of the features using a weight assigned from a support vector machine [19]. Furthermore, Guldogan and Gabbouj [18] tried to utilize standard feature selection algorithms, like IG, to measure a features performance for a given task. Based on these measurements they applied majority voting to produce a ranked list of features further used to select the best working ones. Their evaluation results demonstrate that this method can improve the classification performance and at the same time reduce the computation time.

Currently, we perform a simple feature selection by testing different combinations of features on smaller reference datasets to find the best combinations in terms of processing speed and classification accuracy. For the further system improvement, we will implement several advanced features selection algorithm and will perform a comparison in order to select the best for our use case.

Feature Combination. Features can be combined in two different ways. The first is called feature values fusion or *early fusion*, and it basically fuses values of 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 system implements feature combination using the *late fusion* approach.

Search-based Classification. The search-based algorithm developed in this work has been implemented using *Lire*. Since *Lire* is based on the *Lucene indexes* [16], it also allowed us to create an algorithm that is able to include any type of multimedia data if needed. *Lucene inverted indexes* are created using k-way merge [16]. The index segments are sorted in memory and then merged. Each newly added data element is treated as a new segment and added to existing segments. These indexes have the advantage that they are fast to update and reasonably fast to search. The indexes are field-based and the number of fields is variable depending on the number of used features. The fields are stored using LSH as described before. The algorithm is basically a simple K-NN algorithm, which defines classes c as:

$$c = \arg\max_{\hat{c} \in C} \{ClassScore(\hat{c})\}$$

$ClassScore$ is calculated by summing up the occurrences of each class c and multiplying it with the summed $WeightedRankScore$. $RankScore$ per class is calculated by dividing 1 by the rank for each search query.

$$ClassScore(c) = |c| \sum_{I_i \in \{I_i | Class(I_i) = c\}} RankScore(I_i)^{-1}$$

The $WeightedRankScore$ is the sum of all $RankScore$s in the rank list. 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. The latter is demonstrated by applying it to different use cases described in the following section.

Implementation Details. The indexer is created as a separate tool and in a way that it is easy to distribute over different nodes using, for example, Apache Storm. Indexing is performed when the training data is inserted into the system and is suited for batch processing. Creating the models for the classifier can be done offline and does not influence the real-time capability of the system because it is only done once at the very first time when the training data is inserted into the system. It creates indexes for all directories passed on from the system. The visual features to calculate and store in the indexes can be chosen based on the abnormality because, for different types of diseases, different set of features or combinations are better. For example, bleeding is easier to detect using color features, whereas polyps require also shape and texture information. The indexer stores the generated indexes in a subdirectory inside the indexed directory. If multiple directories are passed for indexing, it creates a separate index for each directory.

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 as described before. 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. We refer to these previously generated indexes, which are searched for similar image features, as classifier indexes or indexes containing training data. The classifier expects at least one classifier index and an input source. The input source can either be a video, an image or another previously generated index. The classifier also includes a benchmarking function that will output the evaluation information and an HTML page with a visual representation of the results, once the processing is finished. The classifier is parallelized and allows to choose how many CPU cores are used to process the data. In the future, a GPU implementation will be supported, because our previous research [44, 46] showed that it can significantly improve the performance.

We have released the source code of the detection subsystem as an open-source project called *OpenSea*[1], under the terms of the GPL version 3[2].

[1] https://bitbucket.org/mpg_projects/opensea
[2] http://www.gnu.org/licenses/gpl-3.0.en.html

3.2.2 Multi-disease Classification. Previously, we claimed that one major difference between our system and related approaches is that it can easily be extended to detect other endoscopic findings (abnormalities, diseases, anatomic landmark or other relevant events during the examination of a patient). To prove that our system is able to perform multi-class classification for diseases beyond polyps, we developed a detection prototype that implements two approaches: global-feature-based and deep-learning-based. Both approaches are tested on a dataset collected from the Bærum Hospital in Norway, one of our collaborators. The amount of data that has been annotated to evaluate the multi-class classification is rather limited so far, and consequently, these results are preliminary.

Multi-class global-feature-based approach (GF-classifier). The basic search-based classification part of the system is used to create a separate classifier for each disease that we want to classify. The difference to the initial version of the detection part is that the ranked lists of each search-based classifier are used in an additional added classification step to determine the final class. For the final classification, we use the random forest classifier (RFC) [7]. It is important to point out that other classification algorithm could be used, and that we choose the random forest approach because it is fast while achieving good results [56].

The RFC creates a forest of classification trees. Each tree is a decision tree that makes, at each of its inner nodes, a branching decision based on one or more feature dimensions. The conditions for these branching decisions are randomly created at the time of the tree's creation, and applied deterministically afterwards. Thus, classes are randomly defined, but features are deterministically classified. To determine the final class, the classifier combines all decisions trees into a final decision using the same late fusion technique used for the features in the standard search-based classifier.

RFC allows parallel classification for each of the separate random trees of the forest. Apart from that the parallel step does also allow for very fast training. Further, the RFC is very efficient for large datasets because of the ability to find distinctive classes in the dataset and also to detect the correlation between these classes. The disadvantage is that training time increases linearly with the number of trees. However, this is not a problem for our use case since training time is not critical. We use the RFC implementation provided by the Weka machine learning library [20].

Multi-class deep-learning-based approach (Deep-classifier). The deep-learning-based classification approach is implemented using Google Tensorflow [1]. As a basis for the deep learning network architecture, we use Inception v3 [52], which is a modern neural network designed for image classification tasks. The Inception v3 model is pre-trained on the ImageNet dataset [13]. From the Inception v3 model, we removed the last layer and retrained it with our medical image classes following the approach presented in [14]. This makes it possible to reuse visual concepts learned from the ImageNet dataset to perform the learning on a smaller dataset.

After removing the final layer from the model, we insert a randomly initialized fully connected layer and retrain the final layer from scratch. All the other layers do not change. This comes with the advantages that not so much training data is needed to train the network, which is a benefit for our medical scenario where lack of good data is a common problem, and that it is faster. Its takes around one day with our settings to retrain the model. The re-trainer is based on an open source implementation [1] of Tensorflow.

At first, we calculate for each image the values for the second last layer (also called bottleneck), which can be seen as kind of features representing the images. These features are then used to retrain the final layer of the network based on the new classes using a softmax function [5]. For the retraining, we run $10,000$ training steps. Each step takes 20 random images in their pre-extracted feature representation to retrain the layer. Because of the small amount of training data, we also perform distortion operations on the images, which is required to avoid network overfitting. In more detail, we perform random cropping, random rescaling and random change of brightness. The grade of distortion is set to 25% per image. In the case of polyp detection, distortions will not destroy the meaning of the image (like it would do if someone, for example, wants to detect letters). After the model has been retrained, it is used as a multi-class classifier that provides the top five classes based on probability for each class.

3.2.3 Localisation Subsystem. The localisation subsystem is intended for finding the exact position of irregularities, which is used to show markers on the disease in the visualization subsystem. All images that we process during the localisation step come from the positive frames list generated by the detection subsystem. The processing of the images is implemented as a sequence of intraframe pre- and main-filters.

Pre-filtering is needed because we use local image features to find the exact position of objects in the frames. Irregularities 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 for example seeds or stool, and lighted by direct and reflected light. Moreover, the image itself can be interleaved, noisy, blurry and over- or under-exposed, and it can contain borders, subimages and a lot of specular reflections (flares) caused by endoscope's light source. Images can have also various resolutions depending on the type of endoscopy equipment used. All these nuances negatively affect the local features detection methods and have to be specially treated to reduce localisation precision impact. In our case, sequence of filters are used to prepare raw input images for the following analysis. These processing steps are border and subimage removal, flare masking and low-pass filtering. After pre-filtering, the images are used for the following local features analysis.

At the moment, we have only implemented localisation of colon polyps using our local feature approach. For future work, we aiming to also localize other irregularities like cancer, bleeding, parasites, etc. The main idea of the localisation algorithm is to use the polyps' 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 round rock connected to an underlying surface with legs varying in thickness. These polyps can be approximated by an elliptically shaped region that consists of local features that differ from the surrounding tissue.

To detect polyps, we use the following sequence of filters: binary noise reduction filter, 2D-gradient filter, threshold border detection filter and binary noise removal filter. The next step creates a filtered binary contour image approximated by a set of ellipses. The

precision of contours approximation via ellipses is measured as distance from ellipses' borders to contours' pixels, which results in an energy map. The final coordinates of one or more polyps in the frame are chosen by looking for maxima in the energy map. For performance reasons, the localiser is implemented in C/C++ and uses *OpenCV* [6]. An example of the output is shown in figure 2.

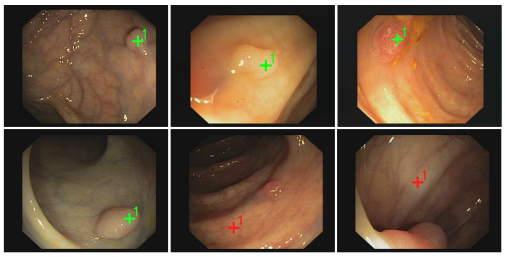


Figure 2: Output of the localisation subsystem marking the possible locations of polyps. The first 4 frames show an exact match, the last two show false positives.

3.3 Visualization and Computer Aided Diagnosis Subsystem

This subsystem has two main purposes. Firstly, it should help in evaluating the performance of the system and get better insights into reason for successes and failures. Secondly, it can be used as a computer-aided diagnosis system for medical experts.

First, we have the *TagAndTrack* subsystem [2] that can be used as a visualisation and computer-aided diagnosis system. Second, we developed an open-source application *ClusterTag* [43] designed for interactive exploration and labeling of big image collections in conjunction with semi-automatic image clustering, annotation and tagging. Third, we developed a web-based visualization that can also be used to support medical experts and is easy to use and distribute. It takes the output of the detection and localisation subsystems and creates a web based visualisation, which later may be combined with a video sharing platform [21, 51], where doctors are able to watch, archive, annotate and share information.

4 SYSTEM EVALUATION

We tested the whole system in terms of accuracy and system performance. For all measurements, we used the same computer (32 cores AMD Opteron 8218 Linux server, 128GB RAM, from 2006). For all experiments, we used the ASU Mayo Clinic polyp database[3]. This is currently the biggest publicly available dataset consisting of 20 videos (converted from WMV to MPEG-4 for the experiments) with a total of 18, 781 frames and different resolution up to full HD (1920x1080) [53].

4.1 Detection and Localisation Accuracy

For detection and localisation accuracy, we used the common metrics, precision, recall and F1 score. All experiments have been conducted on the complete ASU Mayo Clinic polyp database and each subsystem has been evaluated separately.

[3]https://polyp.grand-challenge.org/site/Polyp/AsuMayo/

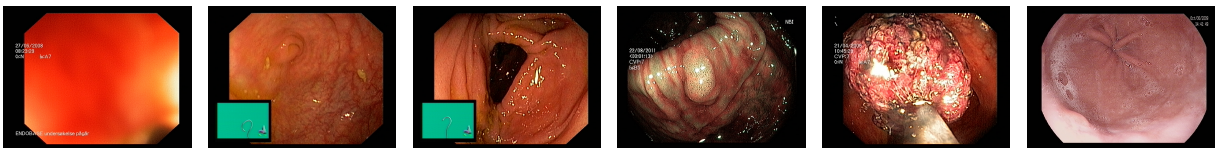
4.1.1 Detection Accuracy. We conducted a leave-one-out cross-validation to evaluate the detection subsystem. This is a method that assesses the generalization of a predictive model. In our case, it describes the process where the training and testing datasets are rotated, leaving out a single different non-overlapping item or portion for testing, and using the remaining items for training. This process is repeated until every item or portion has been used for testing exactly once [15]. Our system allows us to use several different global image features for the classification. The more image features we use, the more computationally expensive the classification becomes. Further, not all image features are equally important or provide equally good results for our purpose. As a first step, we therefore needed to find out which image features we want to use for classification, and we ran the detection with all possible image features in Lire [32] selected on a dataset. Based on this evaluation, feature extractors and descriptors according to Joint Composite Descriptor (JCD) [32] and Tamura [32] (in the following simply called *features* for brevity) were chosen for our measurements due to their promising performance.

To assess the actual performance of the classifier using these two features, we conducted a leave-one-out cross-validation with all available video sequences. With these settings, we achieved an average precision of 0.889, an average recall of 0.964 and an average F1 score value of 0.916. The problem with this average calculation is that different video sequences contribute values based on different numbers of video frames. If we weight the values contributed by every single video sequence with the amount of frames in the sequence, we achieve an average precision of 0.9388, an average recall of 0.9850, and an average F1 score value of 0.9613. In other words, the results mean that we can detect polyps with a precision of almost 94% and we detect almost 99% of all polyp-containing frames. The evaluation is presented in table 1.

Table 1: Performance evaluation by leave-one-out cross-validation for all available videos, using JCD and Tamura features.

Video	True positive	True negative	False positive	False negative	Precision	Recall	F1 score
np_5	1	680	0	0	1	1	1
np_6	1	836	0	0	1	1	1
np_7	1	767	0	0	1	1	1
np_8	1	710	0	0	1	1	1
np_9	1	1,841	0	0	1	1	1
np_10	1	1,923	0	0	1	1	1
np_11	1	1,548	0	0	1	1	1
np_12	1	1,738	0	0	1	1	1
np_13	1	1,800	0	0	1	1	1
np_14	1	1,637	0	0	1	1	1
wp_2	140	9	20	70	0.875	0.6666	0.7567
wp_4	908	1	0	0	1	1	1
wp_24	310	68	127	12	0.7093	0.9627	0.8168
wp_49	421	12	62	4	0.8716	0.9905	0.9273
wp_52	688	101	284	31	0.7078	0.9568	0.8137
wp_61	162	10	165	0	0.4954	1	0.6625
wp_66	223	12	165	16	0.5747	0.9330	0.7113
wp_68	172	51	20	14	0.8958	0.9247	0.9100
wp_69	265	185	138	26	0.6575	0.9106	0.7636
wp_70	379	1	0	29	1	0.9289	0.9631
Weighted average:					0.9388	0.9850	0.9613

4.1.2 Multi-class Classification Accuracy. To evaluate the multi-class classifiers, we collected a new dataset from one of our partner hospitals. The dataset contains six different endoscopic findings that can occur during a colonoscopy with 50 images each, which leads to



(a) Blurry frame (b) Cecum (c) Normal colon mucosa (d) Polyp (e) Tumor (f) Z-line

Figure 3: Example for anatomic findings (classes) in the multi-class dataset.

a total number of 300 images[4]. The classes in the dataset are blurry frames, cecum (pouch that is the beginning of the large intestine), normal colon mucosa (healthy colon wall), polyp, tumor, and Z-line (an anatomic landmark in the colon than can help doctors to orientate). Figure 3 shows one example for each class in the dataset. Because of the small number of images in the dataset, we performed cross-validation. For the cross-validation, we randomly separated the images into 10 different sets of training and test data. 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 even with a smaller number of images, a quite accurate estimation about the performance can be made.

Table 2 shows the confusion matrix (a standard tool for evaluating multi-class classifiers showing the actual class compared to the detected class) for the GF-classifier. The results are a clear indication that this approach performs well. An interesting insight is that normal colon mucosa is often miss-classified as cecum (cecum is also sometimes miss-classified as normal colon mucosa). The example images for cecum (figure 3(b)) and normal colon mucosa (figure 3(c)) reveal that this is not very surprising since it is even hard for a human observer to make a clear decision. Furthermore, from a medical point of view, normal colon mucosa are part of the cecum and under real-world circumstances, this would not be a relevant mistake.

Table 2: Confusion matrix and standard metrics for the six-class classification performance for the multi-class global-features-based approach. The classes are Blurry frames (A), Cecum (B), Normal colon mucosa (C), Polyps (D), Tumor (E), Z-line (F).

		Detected class						*Metrics*		
		A	B	C	D	E	F	Precision	Recall Sensitivity	F1-score
Actual class	A	**250**	0	0	0	0	0	1.0	1.0	1.0
	B	0	**226**	21	3	0	0	0.704	0.904	0.791
	C	0	85	**165**	0	0	0	0.85	0.66	0.743
	D	0	10	8	**226**	6	0	0.953	0.904	0.928
	E	0	0	0	8	**242**	0	0.975	0.968	0.971
	F	0	0	0	0	0	**250**	1.0	1.0	1.0
							Average	**0.914**	**0.906**	**0.91**

The performance of Deep-classifier, which is presented in table 3 can also be considered as good. This approach confuses the classes polyp and cecum more than the GF-classifier, but it is better in detecting normal colon mucosa. For detecting blurry frames and Z-lines, it performs at the same level as the GF-classifier. Based on the confusion matrix for both approaches, we can see that for some classes, the GF-classifier is better and for other classes the Deep-classifier.

[4]The dataset that we could collect in the given time frame with the help of our medical partners is rather small, but it is large enough for a proof-of-concept in combination with cross validation.

Table 3: Confusion matrix and standard metrics for the six-classes detection performance evaluation for the deep-learning-based approach.

		Detected class						*Metrics*		
		A	B	C	D	E	F	Precision	Recall Sensitivity	F1-score
Actual class	A	**250**	0	0	0	0	0	1.0	1.0	1.0
	B	0	**183**	64	3	0	0	0.782	0.732	0.756
	C	0	34	**197**	19	0	0	0.641	0.788	0.707
	D	1	17	45	**183**	4	0	0.875	0.732	0.797
	E	0	0	1	4	**245**	0	0.983	0.98	0.981
	F	0	0	0	0	0	**250**	1.0	1.0	1.0
							Average	**0.879**	**0.872**	**0.876**

Comparison of the GF- and the Deep-classifiers using the standard metrics including precision, recall/sensitivity and F1-score reveals that the GF-classifier outperforms Deep-classifier significantly with a precision of 0, 914, a recall of 0, 906 and a F1-score of 0.91 for the GF-classifier compared to a precision of 0, 879, a recall of 0, 872 and a F1-score of 0.876 for the Deep-classifier.

4.1.3 Localisation Accuracy. Table 4 shows the performance of the localisation subsystem. As ground truth, we used the exact positions of the polyps as provided in the ASU Mayo clinic polyp database. Overall, we reached an average precision of 0.3207, a recall of 0.3183 and an F1 score of 0.3195. The values seem to be rather low, but it is important to point out, that the current localisation algorithm outputs four possible locations per frame. Currently, we are working on an implementation that will be able to output only one location per frame.

Table 4: Performance evaluation of the localisation algorithm in terms of accuracy.

Dataset	True positive	False positive	False negative	Precision	Recall	F1 score
CVC-ClinicDB	397	215	249	0.6487	0.6146	0.6312
ASUMayo 2	1	244	244	0.0041	0.0041	0.0041
ASUMayo 4	443	467	467	0.4868	0.4868	0.4868
ASUMayo 24	74	300	300	0.1979	0.1979	0.1979
ASUMayo 49	36	355	355	0.0921	0.0921	0.0921
ASUMayo 52	194	490	490	0.2836	0.2836	0.2836
ASUMayo 61	129	80	80	0.6172	0.6172	0.6172
ASUMayo 66	92	142	142	0.3932	0.3932	0.3932
ASUMayo 68	63	126	126	0.3333	0.3333	0.3333
ASUMayo 69	0	235	235	0.0000	0.0000	0.0000
ASUMayo 70	4	381	381	0.0104	0.0104	0.0104
Average:				0.3207	0.3183	0.3195

4.2 System Performance

One further requirement for the system is performance. The idea is, as mentioned before, to use the system during live colonoscopies and for mass screening for irregularities in the GI tract, using video sequences, recorded by colonoscopes or WVCs.

For the evaluation, we decided to use the configuration of the system that performed best in the accuracy experiment, because

this scenario will be used in the live system setup, i.e., the global-feature-based version. To enable live assistance for endoscopies, we must reach a frame rate of at least 25 frames per second. For all tests, we used three videos from three different endoscopic devices and different resolutions. The three videos are wp_4 with $1,920x1,080$ and 910 frames, wp_52 with $856x480$ and $1,106$ frames and np_9 with $712x480$ and $1,843$. We chose these three videos because they provide representative examples of the video resolution variations for different types of endoscopic devices.

4.2.1 CPU Processing. For the detection approach, we first measured the indexing part that creates the model that is later on used by the classifier. This process has no real-time requirement and can be seen as batch processing, but it should be feasible for larger datasets. Extracting two features and indexing them for the whole ASU Mayo dataset takes on average 8 milliseconds per frame. There is no big difference between the indexing time of different resolutions. We tested the scaling potential by indexing different datasets. The first dataset $D1$ contains $3,871$ frames, the second one $D2$ contains $14,909$ frames, the third one $D3$ contains $29,818$ frames and the last one $D4$ with $100,000$ frames. Table 5 shows the overall results. We found that a larger dataset leads to a faster indexing time per frame, that is caused by runtime Java code optimizer. Furthermore, we did not find a processing speed increase after more than $30,000$ frames in the dataset. Further processing speed increase is limited by the I/O bottleneck since increasing the number of cores did not increase performance. All in all, our experiments show that the indexer is scalable, can be used with big datasets and it should meet all requirements of the system for future tasks.

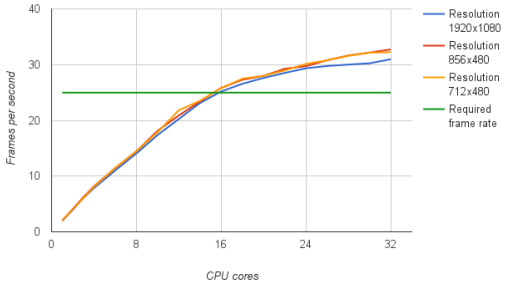
Table 5: Performance evaluation of the indexing part. 4 different datasets with different sizes have been tested to show the scaling capability of the indexing part.

Index	frames	total time in seconds	time per frame in ms
$D1$	3, 871	89.78	23.1
$D2$	14, 909	178.55	11.9
$D3$	29, 818	231.75	7.7
$D4$	100, 000	782.351	7.8

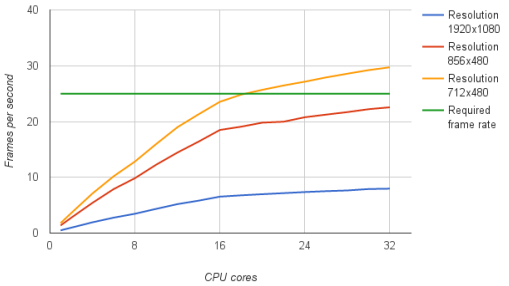
The performance of the detection is more important, since the system should process frames at 25 fps or better to make it usable for live applications. For all tests, we used the 3 different videos described before. Figure 4(a) shows the detection subsystem's performance for the tested videos. The required frames per second for all three resolutions are reached with 16 CPU cores.

Figure 4(b) shows the localisation subsystem's performance for all videos. The required frame rate is not reached for the highest resolution and the best result is 7.9 frames per second. The same is true for the resolution of 856×480. The required frames per second for the lowest resolution are reached with 19 CPU cores used in parallel. The outcome of these experiments clearly shows that our system also can reach real-time requirements for the localisation subsystem but that we need to improve the performance for higher resolutions.

4.2.2 Memory. Figure 5(a) and figure 5(b) show the memory usage for both subsystems. In the localisation, the memory usage behaves normally and shows that the localisation is scalable in terms of memory. For the detection subsystem, the memory usage



(a) The detection subsystem FPS.



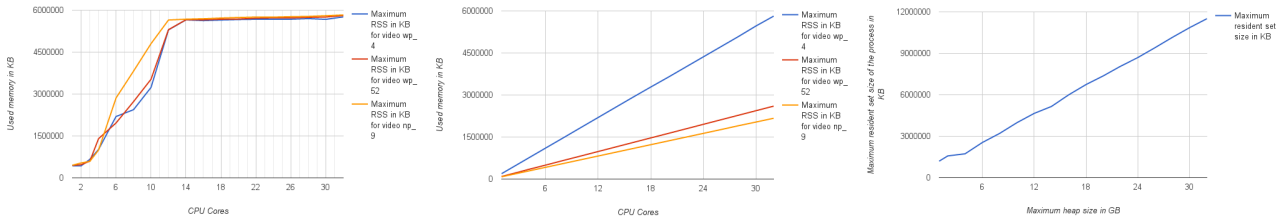
(b) The localisation subsystem FPS.

Figure 4: System performance in terms of frames per second (FPS) depending on the number of CPU cores and the resolution of the videos.

shows an interesting behavior after a certain number of used CPU cores. Therefore, a closer look into it was necessary.

Figure 5(c) depicts this closer look into the detection subsystem memory performance. We tested different memory sizes used for the detection starting from 1GB up to 32GB. This shows that the available memory for the detection part does not influence the frames per second performance. The Java memory scheduler uses as much memory as it can get, but it also performs well with only 1GB. This proves that the detection part does not depend on memory, and therefore, memory is not a bottleneck for scaling.

4.2.3 Size of the Index. A final question that we wanted to answer is if the size of the used classification indexes (number of indexed examples) influences the detection accuracy or system performance. Figure 6 shows the system performance in terms of detection accuracy (F1 score) and frames per second for 3 different training data sizes. The expectation was that smaller indexes would lead to a higher frames per second throughput but with a loss of classification performance. The experiment showed that the index size did not have a significant influence on the number of frames per second output of the detection system. It is possible that an index with several hundred thousand of frames will most probably lead to a lower frames per second output. But, in the intended medical field, a lack of training data is normal. Therefore, this will not influence our system. Another positive aspect is that the classification performance does not decrease with smaller indexes. It is even the opposite, because for wp_52, the F1 score increased slightly compared to the full training data. This shows that the



(a) Memory consumption in the detection part. (b) Memory consumption in the localisation part. (c) Memory consumption versus heap size.

Figure 5: System benchmarks of memory usage.

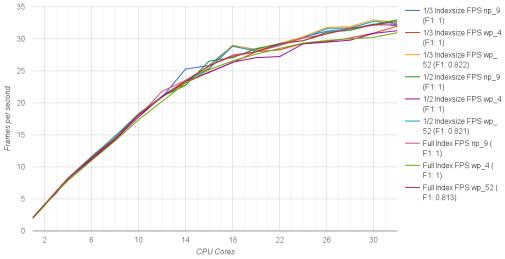


Figure 6: This chart shows how the amount of training data influences the performance of the detection subsystem in terms of frames per second output. The training data has been reduced to 1/2 of the original size (ca. 8, 800 **frames) and 1/3 (ca.** 5, 800 **frames).**

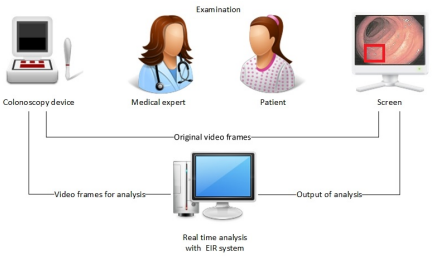


Figure 7: The planned structure of the live system. The medical expert doing a normal examination is assisted in real-time with the results of the video analysis displayed on the auxiliary screen.

detection subsystem also performs very well with a smaller amount of training data matching well our medical scenario.

4.2.4 Initial Cloud Experiments. To investigate what the performance would be on actual hardware for the detection subsystem, some initial tests on Amazon AWS EC2 instances were conducted. On a *c4.8xlarge* instance (Intel Xeon E5-2666-V2 with 36 virtual CPU cores), we were able to classify a video (MPEG-4) with 1, 924 frames and a resolution of $1,920 \times 1,080$ with the features JCD and Tamura, in 29.377 seconds with 65.5 fps. When classifying data from a raw video file the processing time increased to 39.599 seconds with 48.6 fps. When reading the data from a Windows media video (wmv) file, the processing time increased to 40.452 seconds with 47.6 fps. The *c4.8xlarge* instance is the most powerful instance offered by Amazon. We therefore conducted the same tests also on a less powerful *c4.4xlarge* instance (Intel Xeon E5-2666-V2 with 16 virtual CPU cores). Using this instance, we were able to process the MPEG-4 video data in 60.19 seconds with 31.97 fps, the wmv file in 81.17 seconds with 23.7 fps and the raw video file in 79.718 seconds with 24.14 fps. This shows that on newer hardware an even better performance can be achieved.

5 REAL WORLD USE CASES

In this section, we will describe two real world use cases where the presented system can be used. The first one is a live system that will support medical doctors during endoscopies. Currently, we are working on setting it up in one of our partner hospitals. The second one is a system that will automatically analyse videos captured by WVCs. Several hospitals all over Europe and US are involved in this part, and currently, we are collecting data. 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.

5.1 Live System

Figure 7 gives an overview of the proposed live system. Live endoscopy is a common GI examination and is essential for the diagnosis of most mucosal diseases in the gastrointestinal tract, particularly diagnosis of CRC and its precursors. The aim of the live system is to put it between the screen of the doctor and the endoscopy processor. While the endoscopist performs the colonoscopy, the system analyses the video frames that are recorded by the colonoscope. First, we planed to optically show the physician (for example with a red or green frame around the video) when the system detects something abnormal in the actual frame. This can also be extended to determine which disease that the system most probably detected and provide this information to the doctor. Apart from supporting the medical expert during the colonoscopy, the system can also be used to document the procedure. After the colonoscopy, an overview can be given to the doctors where they can make changes or corrections, and add additional information. This can then be stored for later purposes or used in a written endoscopy report. Further, it would be practical to store high quality images of the most important parts. As paper [11] shows, single images can be an efficient way to store important findings from an examination.

5.2 Wireless Video Capsule Endoscope

The present WVCs have a resolution of $256x256$ with 3-10 frames per second (adaptive frame rate with a feedback loop from the receiver to the transmitter). They do not have optimum lighting,

making it difficult use the images. Nevertheless, ongoing work tries to improve the state-of-the-art technology, which will make it possible to use the methods and algorithms developed for colonoscopies also for WVCs [8, 27].

The multi-sensor WVC is swallowed in order to visualize the GI tract for subsequent diagnosis and detection of GI diseases. Thus, people will be able to buy WVCs at the pharmacy, and connect and deliver the video stream from the GI tract to the phone over a wireless network. The video footage can be processed in the phone or delivered to our system, which finally analyses the video automatically. In the best case, the first screening results are available within eight hours after swallowing the WVC, which is the time the camera typically spends traversing the GI tract.

In order to develop such a system, many unsolved tasks need to be addressed through (interdisciplinary) research and development. For example, the training and learning step that allows the system to detect different disease in the GI tract. In the case of the colon, accuracy of existing methods is far below the required precision and recall, and the processing of the algorithms does not scale in terms of big data. Each type of disease or irregularity requires interaction between medical researchers dictating what the system must learn to detect, image processing researchers investigating detection or summarization algorithms, hardware developers to develop/produce/research sensors, distributed processing researchers in order to scale and distribute the (big data) analytics and processing of the sensor data. For other scenarios, like in the upper part of the GI tract, there will be similar challenges and corresponding interaction between research disciplines.

Obviously, the project has high and ambitious goals in developing an end-to-end solution where data recorded by next generation camera and WVCs automatically are processed and algorithmically analyzed for potential pathology in the GI tract. There are large challenges with respect to accuracy (precision and recall), scale of the processing and hardware data quality because of different manufacturers (Olympus and Given are the biggest ones). The aim is to be a leading contributor in the area of medical imaging and sensor processing in the GI tract as well as storing, processing and analysing this type of data. Such next-generation big data applications in the area of medicine are frontiers for innovation and productivity in health systems where there are currently large initiatives both in the EU and the US.

6 CONCLUSION

In this paper, a multimedia system for disease detection and classification in the GI tract has been presented. We briefly described the whole pipeline of the system from annotation (data collection for system learning) to visualisation (doctor feedback). We introduced two new multi-class classification methods, based on global image features and deep learning neural networks. The novelty of the research includes the implementation of a whole system pipeline as a combination of many existing components, as well as several new ones. A detailed evaluation in terms of detection and localisation accuracy and system performance has been performed, and we meet the requirements listed in section 2: (i) high detection accuracy with an F1 score of 96% for polyps, (ii) real-time processing to support live examinations like colonoscopies with a frame rate between 30-65 on the given hardware, (iii) efficient resource utilization to allow massive scale using WVCs shown by both the real-time processing and the low memory consumption, and (iv) expandability to allow the system to support new diseases as shown by the high accuracy multi-disease detection experiment. Our experiments show that the proposed system can achieve equal results to state-of-the-art methods in terms of detection accuracy. Further, we showed that the 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 also presented automatic analysis of WVC videos and live support of colonoscopies as two real world use cases that will benefit from the proposed system and will actually be tested and used in our partner hospitals.

For future work, we plan to improve the detection and localisation accuracy of the system, including even more different abnormalities to detect and work on the localization of irregularities beyond polyps. Presently, we are working with medical experts to collect more training data. As a first result, we just finished two new datasets: an extended multi-class image-dataset for computer aided GI disease detection called Kvasir [42] and a new bowel (colon) preparation quality video dataset called Nerthus [41]. Both datasets are released under open-source and can be used by the community. Additionally, we work on the set-up of the real world use case in the hospitals. Finally, to further improve the performance of the system, we work on an extension that allows the system to use GPUs to further utilize the parallelization potential of the workload.

REFERENCES

[1] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. *Software available from tensorflow. org* 1 (2015).

[2] Zeno Albisser, Michael Riegler, Pål Halvorsen, Jiang Zhou, Carsten Griwodz, Ilangko Balasingham, and Cathal Gurrin. 2015. Expert Driven Semi-supervised Elucidation Tool for Medical Endoscopic Videos. In *Proc. of MMSys*. 73–76.

[3] Luís A Alexandre, Joao Casteleiro, and Nuno Nobreinst. 2007. Polyp detection in endoscopic video using SVMs. In *Proc. of PKDD*. 358–365.

[4] Stefan Ameling, Stephan Wirth, Dietrich Paulus, Gerard Lacey, and Fernando Vilarino. 2009. Texture-based polyp detection in colonoscopy. In *Proc. of BfM*. 346–350.

[5] Christopher M Bishop. 2006. Pattern recognition. *Machine Learning* 128 (2006).

[6] Gary Bradski and Adrian Kaehler. 2008. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Incorporated.

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[8] Rohit Chandra and Ilangko Balasingham. 2015. A microwave imaging-based 3D localization algorithm for an in-body RF source as in wireless capsule endoscopes. In *Proc. of EMBC*. 4093–4096.

[9] Da-Chuan Cheng, Wen-Chien Ting, Yung-Fu Chen, Qin Pu, and Xiaoyi Jiang. 2008. Colorectal polyps detection using texture features and support vector machine. In *Proc. of MDAISM*. 62–72.

[10] Christine Chin and David E Brown. 2000. Learning in science: A comparison of deep and surface approaches. *Journal of Research in Science Teaching* 37, 2 (2000), 109–138.

[11] Thomas de Lange, Stig Larsen, and Lars Aabakken. 2005. Image documentation of endoscopic findings in ulcerative colitis: photographs or video clips? *GE* 61, 6 (2005), 715–720.

[12] R López De Mántaras. 1991. A distance-based attribute selection measure for decision tree induction. *Machine learning* 6, 1 (1991), 81–92.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. of CVPR*. 248–255.

[14] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.. In *Proc. of ICML*. 647–655.

[15] Bradley Efron and Robert Tibshirani. 1997. Improvements on Cross-Validation: The .632+ Bootstrap Method. *J. Amer. Statist. Assoc.* 92, 438 (1997), pp. 548–560.

[16] The Apache Software Foundation. 2013. Apache Lucene - Index File Formats. (2013). https://lucene.apache.org/core/3_0_3/fileformats.html#Definitions Accessed: 2015-07-29.

[17] B. Giritharan, Xiaohui Yuan, Jianguo Liu, B. Buckles, JungHwan Oh, and Shou Jiang Tang. 2008. Bleeding detection from capsule endoscopy videos. In *Proc. of EMBS*. 4780–4783.

[18] Esin Guldogan and Moncef Gabbouj. 2008. Feature selection for content-based image retrieval. *Signal, Image and Video Processing* 2, 3 (2008), 241–250.

[19] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46, 1-3 (2002), 389–422.

[20] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

[21] Pål Halvorsen, Simen Sægrov, Asgeir Mortensen, David KC Kristensen, Alexander Eichhorn, Magnus Stenhaug, Stian Dahl, Håkon Kvale Stensland, Vamsidhar Reddy Gaddam, Carsten Griwodz, and Dag Johansen. 2013. Bagadus: an integrated system for arena sports analytics: a soccer case study. In *Proc. of MMSYS*. 48–59.

[22] Nana Hayashi, Shinji Tanaka, David G Hewett, Tonya R Kaltenbach, Yasushi Sano, Thierry Ponchon, Brian P Saunders, Douglas K Rex, and Roy M Soetikno. 2013. Endoscopic prediction of deep submucosal invasive carcinoma: validation of the narrow-band imaging international colorectal endoscopic (NICE) classification. *Gastrointestinal endoscopy* 78, 4 (2013), 625–632.

[23] Øyvind Holme, Michael Bretthauer, Atle Fretheim, Jan Odgaard-Jensen, and Geir Hoff. 2013. Flexible sigmoidoscopy versus faecal occult blood testing for colorectal cancer screening in asymptomatic individuals. *The Cochrane Library* (2013).

[24] Sae Hwang, JungHwan Oh, W. Tavanapong, J. Wong, and P.C. de Groen. 2007. Polyp Detection in Colonoscopy Video using Elliptical Shape Feature. In *Proc. of ICIP*. 465–468.

[25] Michal F Kaminski, Jaroslaw Regula, Ewa Kraszewska, Marcin Polkowski, Urszula Wojciechowska, Joanna Didkowska, Maria Zwierko, Maciej Rupinski, Marek P Nowacki, and Eugeniusz Butruk. 2010. Quality indicators for colonoscopy and the risk of interval cancer. *New England Journal of Medicine* 362, 19 (2010), 1795–1803.

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

[27] A Khaleghi and I Balasingham. 2015. Wireless communication link for capsule endoscope at 600 MHz. In *Proc. of EMBC*. 4081–4084.

[28] Baopu Li and M.Q.-H. Meng. 2012. Tumor Recognition in Wireless Capsule Endoscopy Images Using Textural Features and SVM-Based Feature Selection. *IEEE Transactions on Information Technology in Biomedicine* 16, 3 (May 2012), 323–329.

[29] Baopu Li and Max Q. H. Meng. 2009. Computer-based Detection of Bleeding and Ulcer in Wireless Capsule Endoscopy Images by Chromaticity Moments. *CBM* 39, 2 (2009), 141–147.

[30] Michael Liedlgruber and Andreas Uhl. 2011. Computer-aided decision support systems for endoscopy in the gastrointestinal tract: a review. *IEEE reviews in biomedical engineering* 4 (2011), 73–88.

[31] Mathias Lux. 2013. LIRE: open source image retrieval in Java. In *Proc. of ACM MM*. 843–846.

[32] Mathias Lux and Oge Marques. 2013. Visual Information Retrieval using Java and LIRE. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 5, 1 (2013), 1–112.

[33] Shawn Mallery and Jacques Van Dam. 2000. Advances in diagnostic and therapeutic endoscopy. *Medical Clinics of North America* 84, 5 (2000), 1059–1083.

[34] A.V. Mamonov, I.N. Figueiredo, P.N. Figueiredo, and Y.-H.R. Tsai. 2014. Automated Polyp Detection in Colon Capsule Endoscopy. *IEEE Transactions on Medical Imaging* 33, 7 (2014), 1488–1502.

[35] Tom M Mitchell. 1997. Machine learning. WCB. (1997).

[36] B. Munzer, K. Schoeffmann, and L. Boszormenyi. 2013. Detection of circular content area in endoscopic videos. In *Proc. of CBMS*. 534–536.

[37] B. Munzer, K. Schoeffmann, and L. Boszormenyi. 2013. Improving encoding efficiency of endoscopic videos by using circle detection based border overlays. In *Proc of ICME workshops*. 1–4.

[38] B. Munzer, K. Schoeffmann, and L. Boszormenyi. 2013. Relevance Segmentation of Laparoscopic Videos. In *Proc. of ISM*. 84–91.

[39] Ruwan Nawarathna, JungHwan Oh, Jayantha Muthukudage, Wallapak Tavanapong, Johnny Wong, Piet C De Groen, and Shou Jiang Tang. 2014. Abnormal image detection in endoscopy videos using a filter bank and local binary patterns. *NC* (2014).

[40] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2014. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv:1412.1897* (2014).

[41] 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, and Pål Halvorsen. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proc. of MMSYS*.

[42] 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, and Pål Halvorsen. 2017. Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proc. of MMSYS*.

[43] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, and Carsten Griwodz. 2017. ClusterTag: Interactive Visualization, Clustering and Tagging Tool for Big Image Collections. In *Proc. of ICMR*.

[44] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Peter Thelin Schmidt, Carsten Griwodz, Dag Johansen, Sigrun L. Eskeland, and Thomas de Lange. 2016. GPU-accelerated Real-time Gastrointestinal Diseases Detection. In *Proc. of CBMS*. 185–190.

[45] Michael Riegler, Martha Larson, Mathias Lux, and Christoph Kofler. 2014. How 'How' Reflects What's What: Content-based Exploitation of How Users Frame Social Images. In *Proc. of MM*. 397–406.

[46] Michael Riegler, Konstantin Pogorelov, Sigrun Losada Eskeland, Peter Thelin Schmidt, Zeno Albisser, Dag Johansen, Carsten Griwodz, Pål Halvorsen, and Thomas de Lange. 2017. From Annotation to Computer Aided Diagnosis: Detailed Evaluation of a Medical Multimedia System. *Transactions on Multimedia Computing, Communications and Applications* 9, 4 (2017).

[47] Michael Riegler, Konstantin Pogorelov, Pŷl Halvorsen, Thomas de Lange, Carsten Griwodz, Peter Thelin Schmidt, Sigrun Losada Eskeland, and Dag Johansen. 2016. EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal endoscopies. In *Proc. of CBMI*. 1–6.

[48] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61 (2015), 85–117.

[49] Jingkuan Song. 2013. Effective hashing for large-scale multimedia search. In *Proc. of SIGMOD/PODS Ph. D. symposium*. 55–60.

[50] Donald F Specht. 1990. Probabilistic neural networks. *Neural Networks* 3, 1 (1990), 109–118.

[51] Håkon Kvale Stensland, Vamsidhar Reddy Gaddam, Marius Tennøe, Espen Helgedagsrud, Mikkel Næss, Henrik Kjus Alstad, Asgeir Mortensen, Ragnar Langseth, Sigurd Ljødal, Ostein Landsverk, Carsten Griwodz, Pål Halvorsen, Magnus Stenhaug, and Dag Johansen. 2014. Bagadus: An Integrated Real-time System for Soccer Analytics. *Transactions on Multimedia Computing, Communications and Applications* 10, 1s, Article 14 (Jan. 2014), 21 pages.

[52] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *arXiv:1512.00567* (2015).

[53] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. 2016. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on Medical Imaging* 35, 2 (2016), 630–644.

[54] Taffee T Tanimoto. 1958. elementary mathematical theory of classification and prediction. (1958).

[55] The New York Times. 2013. The $2.7 Trillion Medical Bill. (2013). http://goo.gl/CuFyFJ Accessed: 2015-11-29.

[56] Brian Van Essen, Chris Macaraeg, Maya Gokhale, and Ryan Prenger. 2012. Accelerating a random forest classifier: Multi-core, GP-GPU, or FPGA?. In *Proc. of FCCM*. 232–239.

[57] L. von Karsa, J. Patnick, and N. Segnan. 2012. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First Edition–Executive summary. *Endoscopy* 44, S 03 (2012), SE1–SE8.

[58] Yi Wang, Wallapak Tavanapong, Johnny Wong, JungHwan Oh, and Piet C de Groen. 2011. Computer-aided detection of retroflexion in colonoscopy. In *Proc. of CBMS*. 1–6.

[59] Yi Wang, Wallapak Tavanapong, Johnny Wong, JungHwan Oh, and Piet C de Groen. 2013. Near Real-Time Retroflexion Detection in Colonoscopy. *BHI* 17, 1 (2013), 143–152.

[60] Yi Wang, Wallapak Tavanapong, Johnson Wong, JungHwan Oh, and Piet C de Groen. 2014. Part-Based Multiderivative Edge Cross-Sectional Profiles for Polyp Detection in Colonoscopy. In *Proc. of BHI*, Vol. 18. 1379–1389.

[61] Yi Wang, Wallapak Tavanapong, Johnny Wong, Jung Hwan Oh, and Piet C De Groen. 2015. Polyp-alert: Near real-time feedback during colonoscopy. *Computer methods and programs in biomedicine* 120, 3 (2015), 164–179.

[62] Yi Wang, Wallapak Tavanapong, Johnny S Wong, JungHwan Oh, and Piet C de Groen. 2010. Detection of quality visualization of appendiceal orifices using local edge cross-section profile features and near pause detection. *BME* 57, 3 (2010), 685–695.

[63] Mingda Zhou, Guanqun Bao, Yishuang Geng, B. Alkandari, and Xiaoxi Li. 2014. Polyp detection and radius measurement in small intestine using video capsule endoscopy. In *Proc. of BMEI*. 237–241.

Paper IX

GPU-accelerated Real-time Gastrointestinal Diseases Detection

GPU-accelerated Real-time Gastrointestinal Diseases Detection

Konstantin Pogorelov[*,•], Michael Riegler[*,•], Pål Halvorsen[*,•], Peter Thelin Schmidt[‡,○]
Carsten Griwodz[*,•], Dag Johansen[♭], Sigrun Losada Eskeland[♣], Thomas de Lange[†,♣]

[*]Simula Research Laboratory, Norway [†]Cancer Registry of Norway [‡]Department of Medicine, Karolinska Institute, Sweden
[•]University of Oslo, Norway [○]Center for Digestive Diseases, Solna and Karolinska University Hospital, Sweden
[♣]Bærum Hospital, Vestre Viken Health Trust, Norway [♭]The Arctic University of Norway, Norway
Email: konstantin@simula.no

***Abstract*—The process of finding diseases and abnormalities during live medical examinations has for a long time depended mostly on the medical personnel, with a limited amount of 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 enables automatic analysis of endoscopy videos, where one use case is live computer-assisted endoscopy that increases 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 ability to perform in real time. The presented system utilizes different parts of a heterogeneous architecture and can be used for automatic analysis of high-definition colonoscopy videos (and a fully automated analysis of video from capsular endoscopy devices). We describe our implementation and report the system performance of our GPU-based processing framework. The experimental results show real-time stream processing and low resource consumption, and a detection precision and recall level at least as good as existing related work.**

***Index Terms*—medical; multimedia; information; systems; classification**

I. Introduction

With the rapid developments in technology that allow miniaturization of cameras and sensors for moving them through the human body, there is an increasing need for real-time medical systems. These improvements lead to a lot of advantages for both patients and doctors, but also challenges for the computer science community. A system supports humans in a critical field like medicine has to fulfill several requirements, including fault tolerance, data security and privacy. Additionally, to support real-time detection of diseases in medical images and videos, the system must exhibit high performance and low resource usage.

In this paper, we describe an new version of system called EIR [1] that provides real-time support for medical image and video data analysis, and we enhance the system with GPU acceleration support. Our goal is to provide an efficient, flexible and scalable analysis and support system for endoscopy of GI tract (see figure 1). It should be applicable both for supporting traditional live endoscopies by giving real-time support and for offline processing of videos generated by wireless capsule endoscopes that are used in large-scale screening. At this time, our system detects abnormalities like those shown in figure 2, in videos of the colon. It does this through a combination of filters using machine learning, image recognition and extraction of global and local image features. However, our system is not limited to this use case, but can be extended to cover analysis of the entire GI tract. Therefore, we developed a live system that can be utilized as a computer-aided diagnostic system and a scalable detection system. In the scenario of medical image processing and computer-aided diagnosis, high precision and recall are important and the object of many studies. Our system must therefore both provide an accurate detection and analysis of the data, and address the often ignored processing performance at the same time. This is important for live feedback during examinations.

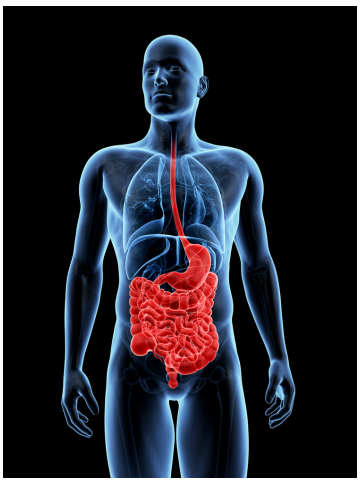


Fig. 1. Our system targets the whole GI tract (Image: kaulitzki/shutterstock.com).

A closer look at the most recent and complete related work, PolypAlert [2], reveals that real-time speeds are not achieved by the current existing systems. To tackle this problem, we have extended and improved the EIR system [3], [4], focusing on the speed of detection. Speedup is gained by applying heterogeneous technologies, in particular graphical processing units (GPUs), where we distribute the workload on a large number of processing cores. The initial results from our experimental evaluation show real-time stream processing and low resource consumption, with a precision and recall of detection at least as good as existing related systems. Compared to existing systems, it is more efficient, scales better with more data at higher resolutions and, it is designed to support different diseases in parallel at run time.

(a) Colon Polyp (b) Colorectal Cancer (c) Ulcerative Colitis (d) Crohn's Disease (e) Diverticulosis

Fig. 2. Some examples of abnormalities that can be found using colonoscopy (images are from Wikimedia Commons).

The rest of the paper is organized as follows. First, we present related work in section II. Then, in section III-A, we briefly describe the base system architecture. This is followed by a presentation of the improved system in section III-B. Next, we present the performance of the system in section IV with polyp detection as a use case. Finally, we draw conclusions in section V.

II. Related Work

Research on automatic detection of abnormalities in the GI tract is usually focused and limited to a very specific disease or abnormality. Most existing work targets detection of polyps in the colon with a specific type of camera, both due to lack of available test data, but also since it is easier to narrow the focus and create more specialized solutions. Systems aimed at polyp detection [5], [6], [7] are promising, but there is a lack of systems that are able to perform their analysis in real-time, which is required to support doctors with computer-aided diagnosis during colonoscopies.

In terms of detection performance, several systems and algorithms have been presented in literature with promising performance. The most recent and also best-performing one is the polyp-detection system of Wang et al. [2]. The presented Polyp-Alert system is able to provide near real-time feedback during colonoscopies. Near real-time in this context is defined as being able to process 10 frames per second. This is done by using visual features and a rule-based classifier to detect the edges of polyps. The system reaches an impressive performance of 97.7% correctly detected polyps. The dataset that has been used for this tests contains 52 videos taken from different colonoscopies. The dataset is not available and a direct comparison is therefore not possible. Polyp-Alert is at the moment limited to polyp detection and does not give real-time feedback for current 25 fps colonoscopy systems.

Nawarathna et al. [8] presented an approach that is not limited to polyp detection in colonoscopy videos. It is also able to detect abnormalities like bleeding. To achieve this, a texton histogram of an image block is used. Nevertheless, this system does not reach real-time performance.

A possible solution to achieve real-time instead of near real-time performance is the SAPPHIRE middleware and software development kit for medical video analysis [9]. The toolkit has been used to built the EM-Automated-RT software [10]. EM-Automated-RT does real-time video analysis to determine the quality of a colonoscopy procedure, and it is able to give visual feedback to the endoscopist performing the procedure. This is done to achieve optimal accuracy of the inspection of the colon during the procedure. Nevertheless, it is limited to the assessment of the endoscopist's quality, and does not automatize disease detection itself.

A dominant trend to speed up processing of CPU-intensive tasks is to offload processing tasks to GPUs. Stanek et al. [9], [10] indicate that utilizing a GPU and program it using either CUDA[1] or OpenCL[2] can be the right way to achieve real-time performance. In other areas this has already been explored to a certain extent. For example, we applied it in sport technology [11], [12], where GPUs were used to improve the video processing performance to achieve live, interactive panning and zooming in panorama video.

In summary, actual computer-aided diagnostic systems for the GI tract do not provide real-time performance in combination with a sufficient detection or localisation accuracy. Therefore, we present a system focusing on both high accuracy detection and real-time performance. Additionally, the aim is to provide flexibility for other diseases that can be detected.

III. System

In our research, we target a general system for automatic analysis of GI tract videos with high detection accuracy, abnormality localisation in the video frames, real-time performance and an architecture that allows easy extensions of the system. In this paper, we focus on achieving real-time performance without sacrificing high detection accuracy.

A. Basic Architecture

Our system consists of three main parts. The first is feature extraction. It is responsible for handling input data such as videos, images and sensor data, and extracting and providing features from it. The most time-consuming aspect here is the extraction of information from the video frames and images.

The second part is the analysis system. Currently, a search-based classifier that is similar to a K-nearest-neighbour approach [13] is implemented. The search-based classifier use more than 20 different global image features and combinations of them for the classification. In our use case of polyp detection, we used an information gain analysis [14] to identify a combination of the features Joint Composite Descriptor (JCD) (which is a combination of Fuzzy Color and Texture

[1]http://www.nvidia.com/object/cuda_home_new.html
[2]http://developer.amd.com/tools-and-sdks/opencl-zone/

Histogram (FCTH) and Color and Edge Directivity Descriptor (CEDD)) and Tamura as the best working ones. The features mainly focus on texture and color, and a detailed description can be found in [15]. Additionally, a localisation algorithm for polyp localisation is supported. The implementation of this part is modular and can be extended with additional diseases, classifiers or algorithms as needed. Of course, adding additional modules will require more computing power to keep the systems real-time ability. We address this by designing a heterogeneous architecture.

The last part is the presentation system. It presents the output of the real-time analysis to the endoscopist. The most challenging aspect here is that the presentation should not introduce any delays, which would make the system unsuitable for live examinations. The presentation of the results is implemented in a light-weight way using web technologies. The advantage is that it does not require additional installations, which sometimes can be problematic in a hospital environment and due to its simplicity it does not consume relevant amounts of resources.

The first version of our system worked on at most two image features at a time, it was restricted to a single computer, and the localisation part did not achieve real-time speed for full high-definition videos. Its performance is given for comparison in section IV-B.

To acheive real-time speed, the architecture had to be improved. We chose to do this by applying heterogeneous processing elements. As discussed in the related work, the most promising approach is the utilization of GPUs.

B. Heterogeneous Architecture Improvement

To improve the performance of our initial basic system architecture, we re-implemented most compute-intensive parts in CUDA. CUDA is a commonly used GPU processing framework for Nvidia graphic cards. We designed an architecture with a heterogeneous processing subsystem as depicted in figure 3.

At the moment, GPU-accelerated processing is implemented for a number of features (JCD, which includes FCTH and CEDD, and Tamura) for the feature descriptor extraction, color space conversion, image resizing and prefiltering.

In our architecture, a main processing application interacts with a modular image-processing subsystem both 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 implementated 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

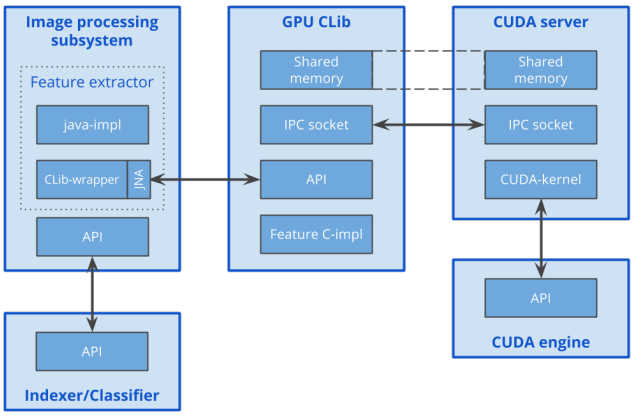


Fig. 3. The main processing application consisting of the indexing and classification parts uses the GPU-accelerated image processing subsystem. This subsystem provides feature extraction and image filtering algorithms. The most compute-intensive procedures are executed on a stand-alone CUDA-enabled 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.

from the CUDA server because they can be integrated more easily with asynchronous on the JNI side then shared-memory semaphores. The CUDA server is implemented in C++ and 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 4. It shows that for the image features, all pixel-related calculations are executed on the GPU. In the 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.

IV. Evaluation

To evaluate our system, we use colorectal polyp detection as a case study. As test data, the ASU-Mayo Clinic polyp database[3] has been used. This dataset is the largest publicly available dataset consisting of 20 videos. We converted the videos from WMV to MPEG-4 for the experiments. The 20 videos have a total number of 18.781 frames with a maximum resolution of 1920×1080 pixels (full high definition) [16]. Further, we concentrate the experiment on the detection part.

[3]http://polyp.grand-challenge.org/site/Polyp/AsuMayo/

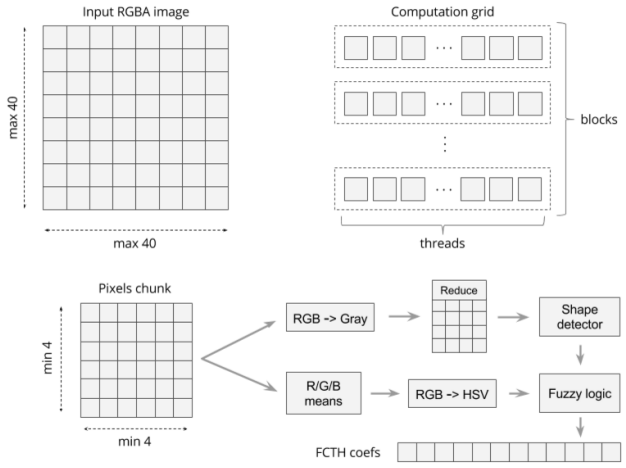


Fig. 4. GPU-acceleration is used to extract various features from input frames. The figure shows an example of our FCTH feature implementation. The input frame is split into a number of non-overlapping blocks. Each of them is processed separately by two GPU-threads. The main processing steps include color space conversion, size reduction, shape detection and fuzzy logic computations.

TABLE I
LEAVE-ONE-OUT CROSS-VALIDATION FOR 20 VIDEOS IN THE USED DATASET. THE TABLE DEPICTS TP (TRUE POSITIVES), TN (TRUE NEGATIVES), FP (FALSE POSITIVES), FN (FALSE NEGATIVES) AND THE METRICS PRECISION, RECALL AND F1 SCORE.

Video	TP	TN	FP	FN	Precision	Recall	F1
np_5	1	680	0	0	1	1	1
np_6	1	836	0	0	1	1	1
np_7	1	767	0	0	1	1	1
np_8	1	710	0	0	1	1	1
np_9	1	1,841	0	0	1	1	1
np_10	1	1,923	0	0	1	1	1
np_11	1	1,548	0	0	1	1	1
np_12	1	1,738	0	0	1	1	1
np_13	1	1,800	0	0	1	1	1
np_14	1	1,637	0	0	1	1	1
wp_2	140	9	20	70	0.875	0.6666	0.7567
wp_4	908	1	0	0	1	1	1
wp_24	310	68	127	12	0.7093	0.9627	0.8168
wp_49	421	12	62	4	0.8716	0.9905	0.9273
wp_52	688	101	284	31	0.7078	0.9568	0.8137
wp_61	162	10	165	0	0.4954	1	0.6625
wp_66	223	12	165	16	0.5747	0.9330	0.7113
wp_68	172	51	20	14	0.8958	0.9247	0.9100
wp_69	265	185	138	26	0.6575	0.9106	0.7636
wp_70	379	1	0	29	1	0.9289	0.9631
Weighted average:					0.9388	0.9850	0.9613

Localisation of the polyp in the frame is also implemented and optimized, but due to space restrictions, it is not included here.

A. Polyp Detection

In terms of detection performance, we reach acceptable results, as illustrated in table I. The actual performance of the system has been assessed using a combination of JCD and Tamura features. For a robust and representative evaluation, we conducted a leave-one-out cross-validation with all available video sequences. The training of the system using 19 videos takes around 2 minutes. Due to the problem that different video sequences contribute values based on different numbers of video frames, we weighted the values contributed by every single video sequence with the overall number of frames in the sequence. This led to an average precision of 0.9388, an average recall of 0.9850, and an average F1 score value of 0.9613. That means that the system can find polyps with a precision of almost 94% and detect almost 99% of all frames that contain a polyp.

These results demonstrate that the system is able to reach high detection accuracy and also, that it can compete with other state-of-the-art systems. For example, Wang et al. [2] reach with their system a recall of 97.70% while our system reaches 98.50%. Hwang et al. [17] report a precision of 83.00% while we achieve 93.88%. In terms of sensitivity, we reach 96.37% compared to Wang et al. [18] with 81.40%, Alexandre et al. [19] with 96.69% and Cheng et al. [20] with 86.20%. Thus, our system performs at the high level of precision compared to the best related systems. However, more important in this paper is the comparison of our own basic architecture with the improve heterogeneous approach in terms of their time-performance.

B. Live Analysis in Real-time

Basic Architecture. The basic multi-core CPU-only architecture performance results are depicted in figure 5. For all the tests, we used 3 videos from 3 different endoscopic devices and different resolutions. The three videos are wp_4 with $1,920\times1,080$, wp_52 with 856×480 and np_9 with 712×480. We chose these videos to show the performance under the different requirements that the system will have to face when in practical use. The computer used was a Linux server with 32 AMD CPUs and 128 GB memory. The figures show, that the basic 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 multi-CPU systems. In terms of memory, tests showed that the system has rather small requirement. This is good, since it means that memory consumption is not a bottleneck to scalability, and that we can ignore it for now.

Heterogeneous 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 labelled correspondingly in figures 6, 7, 8 and 9. 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

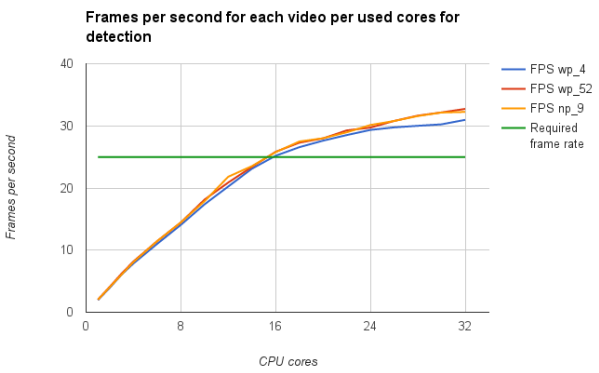


Fig. 5. The detection performs efficiently and the required frame rate is reached with 12 GB of memory and 16 CPU cores used in parallel on cluster-based computation platform without utilizing heterogeneous architecture.

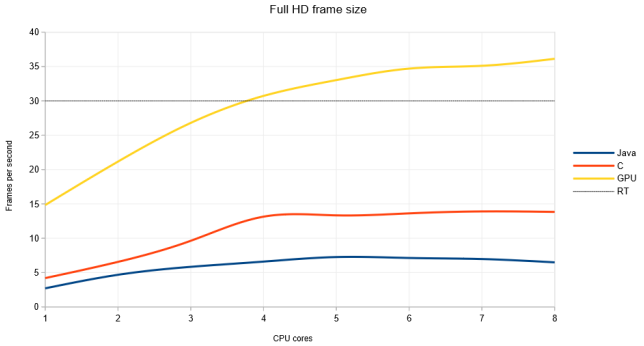


Fig. 6. The improved GPU-enabled heterogeneous algorithm reaches real-time performance (RT line) with 30 frames per second for full HD (1920×1080) videos on a desktop PC using only 4 CPU cores and 5 Gb of memory. The maximum frame rate is around 36 FPS using 8 CPU cores. The Java and C implementations cannot reach real-time performance on the used hardware.

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 basic 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 6) with corresponding frame processing times of 154ms and 72ms, respectively (see figure 8). For the smaller frame sizes, real-time speed was reached at most 4 CPU cores and at most 4 GB of memory. The maximum frame rates that were be reached were 49 FPS, 51 FPS and 66 FPS for WVGA1, WVGA2 and CIF frame sizes, respectively (see figure 7 and figure 9).

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 6) and up to 5 Gb of memory with a frame processing time of 32.6ms (see figure 8). 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 7 and figure 9).

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

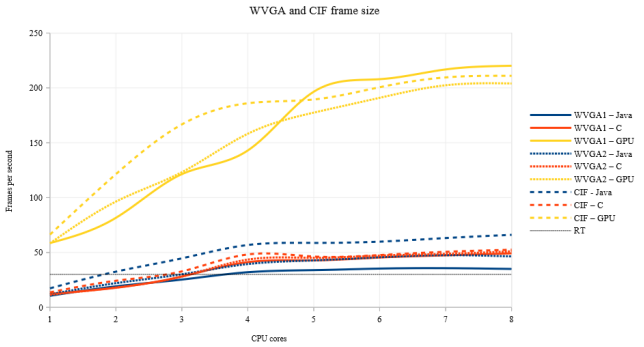


Fig. 7. The smaller WVGA1 (856×480), WVGA2 (712×480) and CIF (384×288) videos can be processed by the improved GPU-enabled heterogeneous algorithm in real-time using only 1 CPU core. The maximum frame processing rate reaches more than 200 FPS. These results can be improved by putting all feature-related computations on the GPU.

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.

V. Conclusion

Efficient and fast data analysis of medical video data is important for to several reasons, including real-time feedback and increased system scalability. In this paper, we have presented a computer-based medical systems that tackles live automatic analysis of endoscopy videos. The presented system utilizes different parts of heterogeneous architectures and will soon be tested in a clinical trial with high definition colonoscopy videos. Compared to existing systems, our system provides an abnormality detection precision and recall level at least as good as existing related work. However, with an achieved

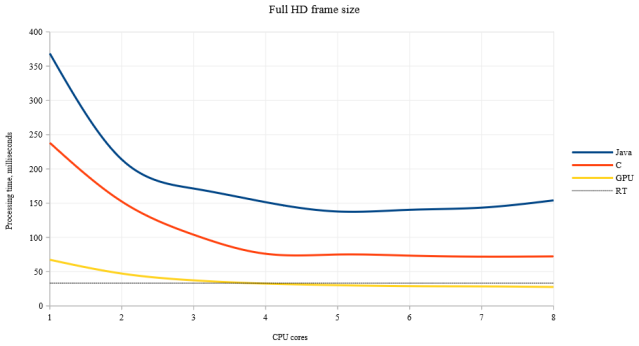


Fig. 8. The processing time for the GPU-accelerated algorithm decreases slightly with increasing number of used CPU cores for a single full HD frame. This happens due to the CPU-parallel implementation of feature comparison and search algorithms which are not as compute intensive as feature extraction. 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.

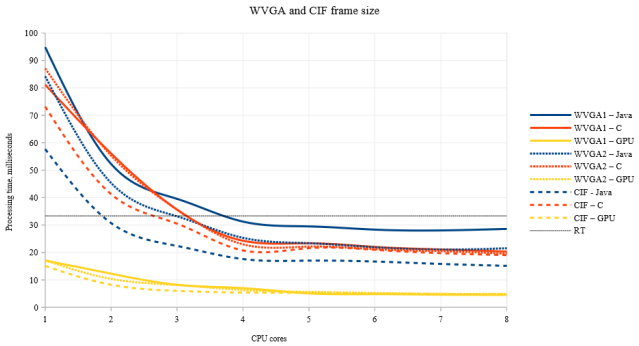


Fig. 9. For the smaller frame sizes the GPU-accelerated algorithm results in a processing time far below the real-time margin. The minimum is reached with 5 milliseconds using 8 CPU cores. This is a prove for the high system performance and ability to be extended by additional features or to process several video streams at the same time on a conventional desktop PC.

performance of 200 frames per seconds, it is superior with respect to video stream processing time and the ability to provide real-time automatic feedback during live endoscopies.

We continue to optimize and improve our implementation of the detection system. Ongoing work includes moving the localisation to the GPU, and we are in the process of extending the number of diseases detected. Our current performance easily allows for this, and our future multi-disease detection system will be distributed on several computers.

Acknowledgment

This work has been founded by the Norwegian Research Council under the FRINATEK program, project "EONS" (#231687).

References

[1] 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 *Proc. of CBMI*, 2016.

[2] 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*, no. 3, 2015.

[3] M. Riegler, K. Pogorelov, J. Markussen, M. Lux, H. K. Stensland, T. de Lange, C. Griwodz, P. Halvorsen, D. Johansen, P. T. Schmidt, and S. L. Eskeland, "Computer aided disease detection system for gastrointestinal examinations," in *Proc. of MMSys*, 2016.

[4] K. Pogorelov, M. Riegler, J. Markussen, H. Kvale 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 *Proc. of MMSys*, 2016.

[5] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Near real-time retroflexion detection in colonoscopy," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 143–152, 2013.

[6] Y. Wang, W. Tavanapong, J. S. Wong, J. Oh, and P. C. de Groen, "Detection of quality visualization of appendiceal orifices using local edge cross-section profile features and near pause detection," *IEEE Biomedical Engineering (BME)*, vol. 57, no. 3, pp. 685–695, 2010.

[7] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Computer-aided detection of retroflexion in colonoscopy," in *Proc. of IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 2011, pp. 1–6.

[8] R. Nawarathna, J. Oh, J. Muthukudage, W. Tavanapong, J. Wong, P. C. De Groen, and S. J. Tang, "Abnormal image detection in endoscopy videos using a filter bank and local binary patterns," *NC*, 2014.

[9] S. R. Stanek, W. Tavanapong, J. Wong, J. Oh, R. D. Nawarathna, J. Muthukudage, and P. C. De Groen, "Sapphire middleware and software development kit for medical video analysis," in *Proc. of CBMS*, 2011, pp. 1–6.

[10] ——, "Sapphire: A toolkit for building efficient stream programs for medical video analysis," *Computer methods and programs in biomedicine*, vol. 112, no. 3, pp. 407–421, 2013.

[11] H. K. Stensland, V. R. Gaddam, M. Tennøe, E. Helgedagsrud, M. Næss, H. K. Alstad, A. Mortensen, R. Langseth, S. Ljødal, Ø. Landsverk *et al.*, "Bagadus: An integrated real-time system for soccer analytics," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 10, no. 1s, p. 14, 2014.

[12] R. Langseth, V. R. Gaddam, H. K. Stensland, C. Griwodz, and P. Halvorsen, "An evaluation of debayering algorithms on gpu for real-time panoramic video recording," in *Proc. of ISM*, 2014, pp. 110–115. [Online]. Available: http://dx.doi.org/10.1109/ISM.2014.59

[13] 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 *CBMI*, 2016.

[14] J. T. Kent, "Information gain and a general measure of correlation," *Biometrika*, vol. 70, no. 1, pp. 163–173, 1983.

[15] M. Lux and O. Marques, *Visual Information Retrieval Using Java and LIRE*. Morgan & Claypool, 2013, vol. 25.

[16] N. Tajbakhsh, S. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, 2015.

[17] 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*, Sept 2007, pp. 465–468.

[18] 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," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1379–1389, 2014.

[19] L. A. Alexandre, J. Casteleiro, and N. Nobreinst, "Polyp detection in endoscopic video using svms," in *Proc. of PKDD*, 2007, pp. 358–365.

[20] 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*. Springer, 2008, pp. 62–72.

Paper X

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

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

Konstantin Pogorelov[1], Michael Riegler[1], Jonas Markussen[1], Håkon Kvale Stensland[1]
Pål Halvorsen[1], Carsten Griwodz[1], Sigrun Losada Eskeland[3], Thomas de Lange[23]

[1]Simula Research Laboratory and University of Oslo
[2]Cancer Registry of Norwayy [3]Vestre Viken Hospital Trust

konstantin@simula.no

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.

CCS Concepts

•**Information systems** → **Information retrieval; Multimedia and multimodal retrieval;**

Keywords

Medical Multimedia; Information Systems; Classification

1. INTRODUCTION

Colonoscopy is a medical procedure, during which specialists in bowel diseases (gastroenterologists), investigate and operate on the colon through minimally invasive surgery by using flexible endoscopes. These examinations are usually done in a special examination room as depicted in figure 1(a). A standard hospital normally has several of these rooms in their gastroenterology department. These rooms contain screens for the doctors that show the video stream from the camera, a bed for the patient, the endoscopic processor, a desktop computer for reporting and some medical treatment supplies. The endoscopes can vary in their attributes like the thickness of the endoscope or its length, but also in the resolution of the videos. Figure 1(b) shows a collection of different endoscopes. Endoscopes are frequently moved between examination rooms to fit the requirements of a specific examination. From the tip of the endoscope (figure 1(c)), a video is transmitted, and the gastroenterologist relies on the video stream to diagnose disease and apply treatments. To control the endoscope, the control unit that is part of every endoscope is used. As one can see in figure 1(d), this is a complex mechanism that requires a lot of concentration from the doctor during the whole procedure, lasting up to 2 hours depending on the findings. The camera can be seen as the virtual the eye of the gatroentologists, and the video stream is all they perceive. Usually, doctors get "third eye" support from their nurses to support them during the examinations and increase the number of findings.

Recently, computer-aided diagnostic systems are more and more used in gastroenterology. The most recent and best

(a) The examination room where the endoscopies are performed. A usuall hospital has several of these rooms.

(b) Different endoscopes for different examinations and patients. For example the very small one is for children.

(c) The tip of the endoscope. It is very flexible and can be moved by the gastroentologist in every possible direction.

(d) The controll unit of the endoscope the gastroentologiest uses to controll the endoscope in terms of zoom, rotation, etc.

Figure 1: These images show an auditorium and endoscopic equipment in the Bærum Hospital in Norway where our system will be used.

MMSys'16 May 10-13, 2016, Klagenfurt, Austria

ACM ISBN 978-1-4503-4297-1/16/05.

DOI: http://dx.doi.org/10.1145/2910017.2910636

working system is Polyp-Alert [10]. This computer-aided diagnostic system helps to determine the quality of the colonoscopy during the procedure. It reaches very high accuracy and sensitivity, but it only reaches near real-time and not full real-time feedback. This is not optimal for live examinations where the medical expert controls the camera manually and cannot rely on a system that introduces delays. Even though real-time performance 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 and to use Device Lending whenever more resources are needed. Here, Device Lending is a concept where computers interconnected in a PCI Express network can share devices using a transparent cross-machine device sharing system without any special effors to use remote resources locally. It is a low-latency, high-throughput solution for distributed computing, utilizing common hardware already present in all modern computers and requiring little additional interconnection hardware.

In this paper, we will present a demo that utilizes Device Lending of GPUs in combination with our own computer-aided diagnosis system. With this demo, we address two main challenges. First, we will show that real-time support is possible using this technology. Second, we demonstrate the possibility of having one mainframe that can lend the devices to different computers based on the computational demands. This can be an important advantage and even required for scenarios where no room for large machines exists. Further, it can be important for setups where the requirements change fast and often on the fly (e.g., an examination room in a hospital changes the used endoscopes several times during the day; endoscopes with a very high resolution need more processing power than those with lower resolution).

2. REAL-TIME COMPUTER AIDED DIAGNOSIS SUPPORT

Automatic detection of polyps in colonoscopies has been in focus of research for a long time [9]. However, few complete systems exist that are able to do real-time detection, or that can support endoscopists by computer-aided diagnosis for colonoscopies in real-time and at the same time maintain a high detection accuracy. The most recent and best working approach is Polyp-Alert [10] that is able to give near real-time feedback during colonoscopies. Visual features and a rule based classifier are used to detect the edges of polyps, and a performance of 97.7% correctly detected polyps is reported. However, real-time support is limited as they reach only 10 frames per second.

To target the real-time performance, we have proposed EIR [8, 7, 6] medical experts supporting system for the task of detecting diseases and anatomical landmarks in the gastrointestinal (GI) tract, which used in this demo as a use case. It has several key attributes, i.e., EIR (i) is easy to use, (ii) is easy to extend to different diseases, (iii) can do real time handling of multimedia content, (iv) is able to be used as a live system and (v) has high classification performance with minimal false negative classification results. Compared to Polyp-Alert, our detection accuracy is slightly below. The classification performance of the polyp detection in our EIR system lies around a precision of 0.903 and a recall of 0.919, but it is tested on a different dataset, meaning that the numbers are not directly comparable.

Currently, the system consists of two parts, the detection subsystem that detects irregularities in video frames and images and the localisation subsystem that localises the exact position of the disease. The detection can not determine the location of the found irregularity. The location determination is done by the localisation subsystem. The localisation subsystem uses the output of the detection system as input. After the automatic detection and analysis of the content, the output has to be presented in a meaningful way to the gastroentologists. Therefore, the system has a visualisation subsystem that is reliable, robust and easy to understand also under stressful situations that can occur during a live examination. Moreover, it supports easy search and browsing through a large amount of data after the examination. In this demo, we do not focus on EIR but rather using Device Lending and how it can improve performance. EIR itself is just a relevant use case.

2.1 GPU Implementation

Parts of EIR had to be improved and changed to run on multiple GPUs and allow the system to perform in real-time. Therefore, the most compute-intensive parts have been ported to CUDA, a computation support framework for nVidia graphic cards. To achieve this, parts of the system had to be built as a heterogeneous processing subsystem. The GPU framework supports at the moment a number of features, namely Joint Composite Descriptor (JCD), which includes Fuzzy Color and Texture Histogram (FCTH) and Color and Edge Directivity Descriptor (CEDD), and Tamura, but we are working on increasing the supported features.

A main processing application interacts with a modular image processing subsystem. Both of these are implemented in Java. A multi-threading architecture is used by the image processing unit to handle multiple processing and feature extraction requests at the same time. A shared library that is responsible for maintaining connection with and stream data to the stand-alone CUDA-enabled processing server is implemented in C++. To ensure high data transfer performance and reduce excessive data copy operations, shared memory has been used, while sending requests and receiving status responses uses local UNIX sockets. A CUDA server implemented in C++ runs in the background and performs computations on GPU. The whole system can easily be extended with multiple CUDA servers running locally or on a number of remote servers. This is also valid for the processing server, which can be extended with new feature extractors and advanced image processing algorithms, and utilize multi-core CPU and GPU resources concurrently.

2.2 Device Lending

Device Lending is a concept where computers interconnected in a PCI Express [5] network can share devices. It provides transparent, low-latency cross-machine PCIe device sharing without *any* need to implement application-specific distribution mechanisms or modify native device drivers. As the workload increases or decreases, the system can allocate and de-allocate additional resources.

Today, PCIe is the most common interconnection network inside a computer, and with PCIe non-transparent bridges (NTB) [1], it can be turned into an interconnection network

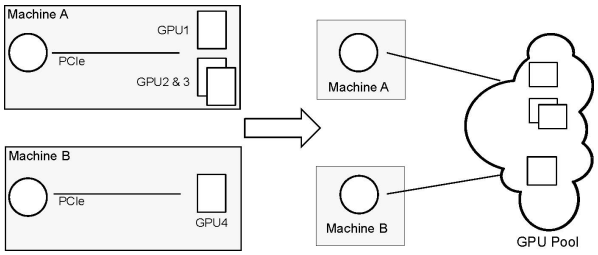


Figure 2: Pooling of devices attached in the PCIe network in the experimental setup.

for multiple machines. In PCIe, all devices connected to the computer are considered part of one common resource pool (figure 2). All devices resources in PCIe are represented by addresses that can be mapped into a remote memory space by an NTB. Device Lending is implemented [3] using Dolphin Interconnect Solutions NTB software [1].

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 [2] or MPI [4].

2.3 Performance Evaluation

To evaluate the performance 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 performed experiments can be found in table 1. 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 2), 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. With the current machine setup it is not possible to lend more that one GPU because of software limitations in the motherboard's BIOS.

In the experiments, 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 performance from capturing the video up to showing the output on the screen. The complete evaluation is shown in figure 3.

Figure 3(a) shows the performance in terms of processing time per frame for all streams simultaneous. The results

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

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

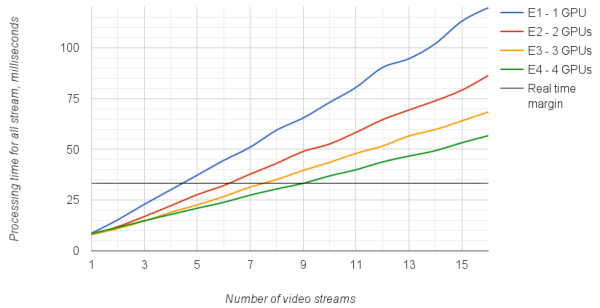
reveal that for up to 7 parallel full HD streams, the 3 local GPUs are fast enough. For more than 7 streams, GPU lending is required. The graph shows that the more parallel streams are processed, the better is the performance gain from the borrowed GPU. We assume that this is due to the excessive overhead 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, and that Device Lending can indeed improve performance.

The plot in figure 3(b) shows the overall system performance. The evaluation shows that Device Lending can indeed improve the 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. This is another indicator that Device Lending can increase performance a lot for large scale processing.

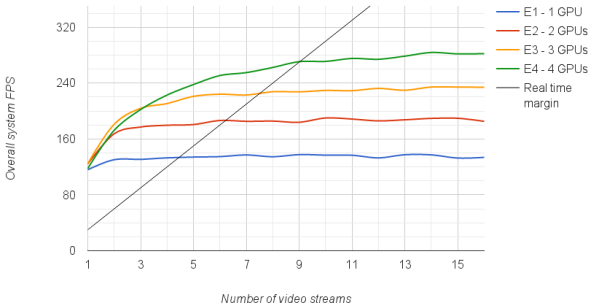
All in all, the experiments showed two important things: (i) Device Lending does not make sense for small amounts of data, but if the data to process is large it can give a large performance boost, and (ii) Device Lending makes sense in a multi-auditory scenario like we present with our demo.

3. DEMONSTRATION SETUP

The above experiments show the performance of EIR on powerful machines and that Device Lending works efficiently, i.e., high performance and low latencies at a very low overhead. However, placing such a setup in the many examination rooms in a hospital is impractical for a number of reasons like high costs and noisy machines. A more realistic scenario is therefore to have smaller machines in each room and use Device Lending whenever more resources are needed.



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



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

Figure 3: System performance evaluation in terms of processing time per frame and maximum performance using 4 different configurations described in table 1. Each video stream is a full HD video.

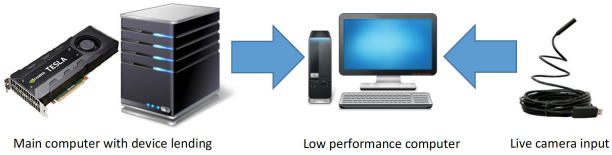


Figure 4: A compete overview of the demo setup. The demo consists of 2 computers, 1 Dolphin interconnect device, 1 screen, an artificial colon and a flexible camera. The users can use the camera in the flexible colon and will get real-time feedback about possible findings. Furthermore, the demo can be switched between Device Lending on and off to demonstrate the effect of it more clear.

To demonstrate the usefulness of Device Lending, we therefore use the above scenario. In the demo, users can use a flexible camera to perform a colonoscopy in an artificial colon, and the system will support them in real-time with analysis and feedback. The complete demo setup is depicted in figure 4. During the demo, the camera can be used to examine the artificial colon and the output of the system will be shown in real-time on the screen. The demo will show the performance increase when a GPU can is borrowed from another machine. Therefore, the demo application can be switched between lending and not lending a GPU. An example of the output for detected polyps can be seen in figure 5. This setup is similar to our real world setup of the system for live colonoscopy with videos as shown to the doctors. Thus, the processing will be done on a very weak computer that is not able to perform the complicated analysis in real-time. Therefore, it is connected to another PC via a Dolphin interconnect device and uses Device Lending to allocate the required processing power. The demo will clearly show the visible differences when Device Lending is used and when not. We also would like to point out, that the presented demonstration is based on the findings in [3] which describes the Device Lending in more detail for further reading.

4. CONCLUSION AND FUTURE WORK

In this paper, we presented a demo for Device Lending for computer-aided diagnosis that can assist medical doctors to analyse colonoscopy videos in a multi-auditory scenario. We proved that we can reach high performance in terms of processing time for several full HD video streams in parallel which make it possible to use the proposed system during several and parallel live colonoscopies. We showed that running multiple classifiers in parallel by offloading the processing to multiple machines connected through a PCI Express network and using GPU lending works in our scenario. This optimized version of the application will be able to dynamically allocate, distribute and release compute resources on demand from a pool of available GPUs. For future work, we would like to improve the scheduling of tasks within our lending network. This would include decisions for what and how much to lend to which part of the system using different input information like the required support level of doctors and the endoscope used. We also think that this idea is applicable to other scenarios like for example in cinemas where a less powerful PC in each saloon allocates GPUs based on the quality of the movie to show, e.g., one room shows 4k, one 3D and another one full HD.

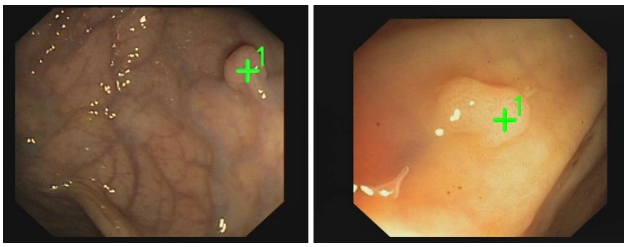


Figure 5: This figure shows 2 examples of what the doctor will see on the screen and what we will show during the demo. In both pictures, the system detected polyps and marked them with a cross. If nothing is detected, the corners of the screen are marked green for feedback.

5. ACKNOWLEDGMENT

This work has been performed in context of the FRINATEK project *EONS* (#231687) and the BIA project *PCIe* (#235530) funded by the Research Council of Norway (RCN). The authors also acknowledge Lars Bjørlykke Kristiansen and Dolphin Interconnect Solutions for assistance with Device Lending and PCIe interconnect equipment. We also would like to thank Mathias Lux from the University of Klagenfurt for "lending" us hardware at the conference venue.

6. REFERENCES

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

[2] 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.

[3] 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.

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

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

[6] 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 *Proc. of CBMS*, 2016.

[7] 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 *Proc. of CBMI*, 2016.

[8] M. Riegler, K. Pogorelov, J. Markussen, M. Lux, H. K. Stensland, T. de Lange, C. Griwodz, P. Halvorsen, D. Johansen, P. T. Schmidt, and S. L. Eskeland. Computer aided disease detection system for gastrointestinal examinations. In *Proc. of MMSys*, 2016.

[9] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Near real-time retroflexion detection in colonoscopy. *IEEE BMHI*, 17(1):143–152, 2013.

[10] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. de Groen. Polyp-alert: Near real-time feedback during colonoscopy. *CMPBM*, 120(3):164–179, 2015.

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

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

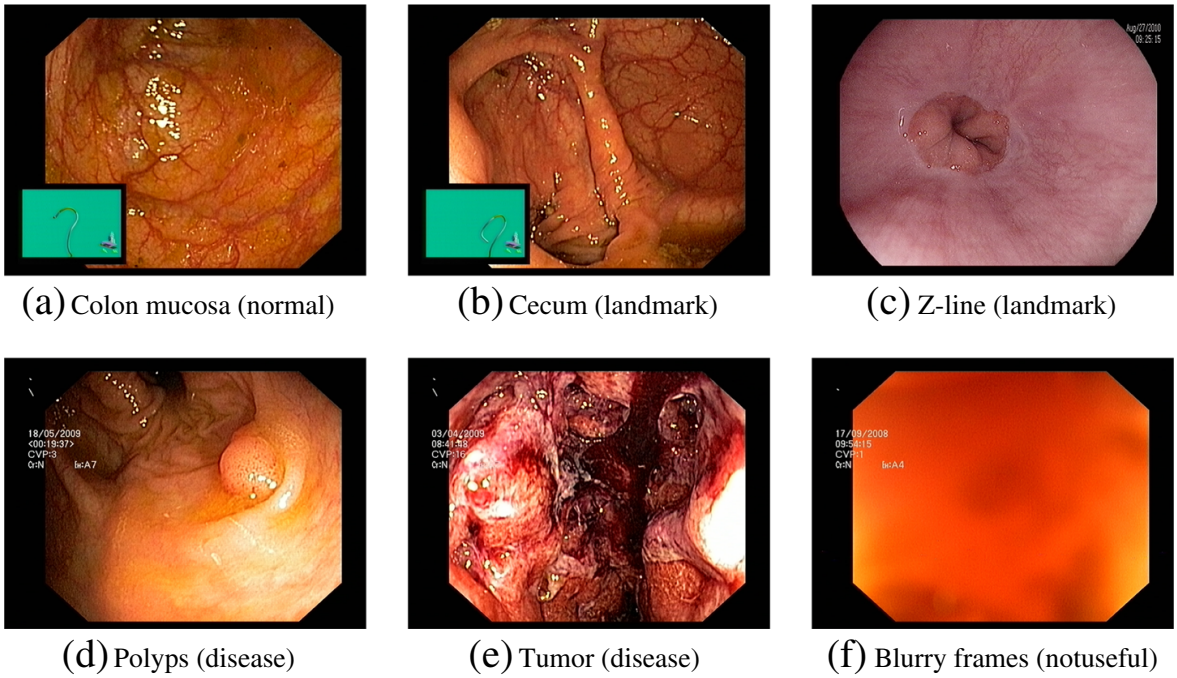


Fig. 1 Example frames from human colon showing normal tissue (**a**)–(**c**), abnormal findings (**d**)–(**e**) and useless frames (**f**)

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/

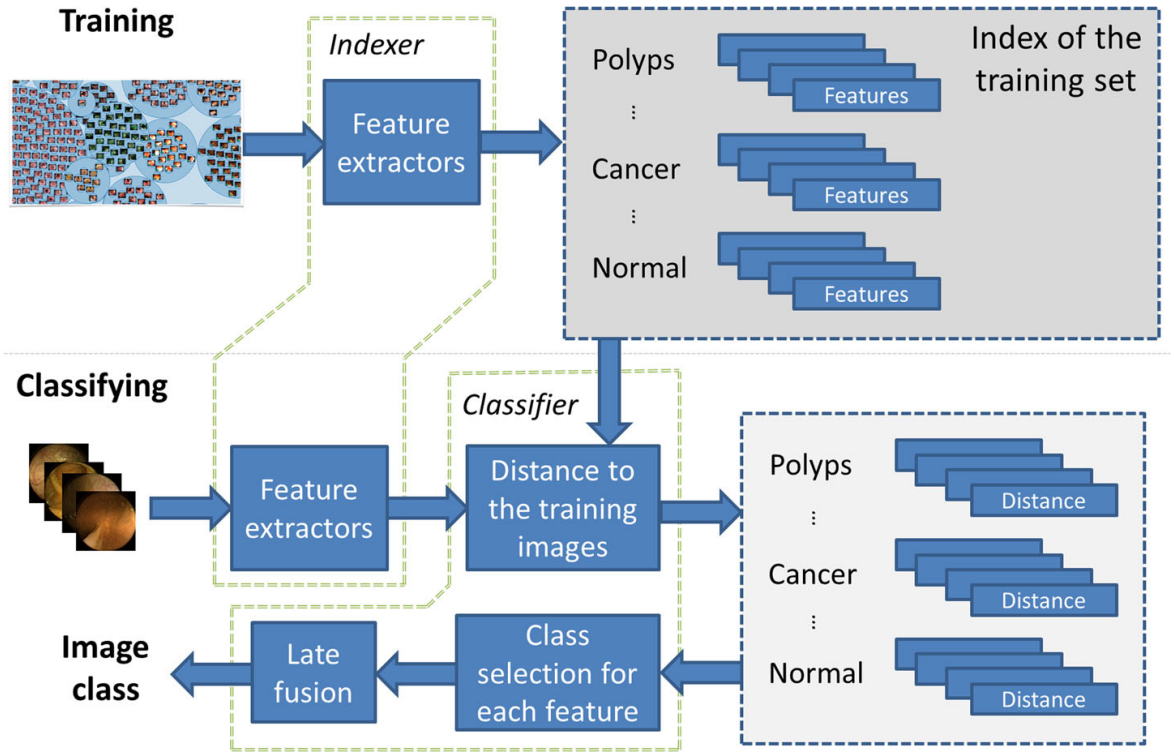


Fig. 2 Detailed steps for the multi-class global-feature-based detection implementation

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

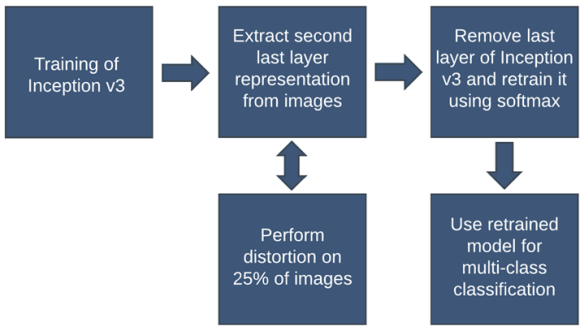


Fig. 3 Detailed steps for the neural network approach based detection implementation

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/

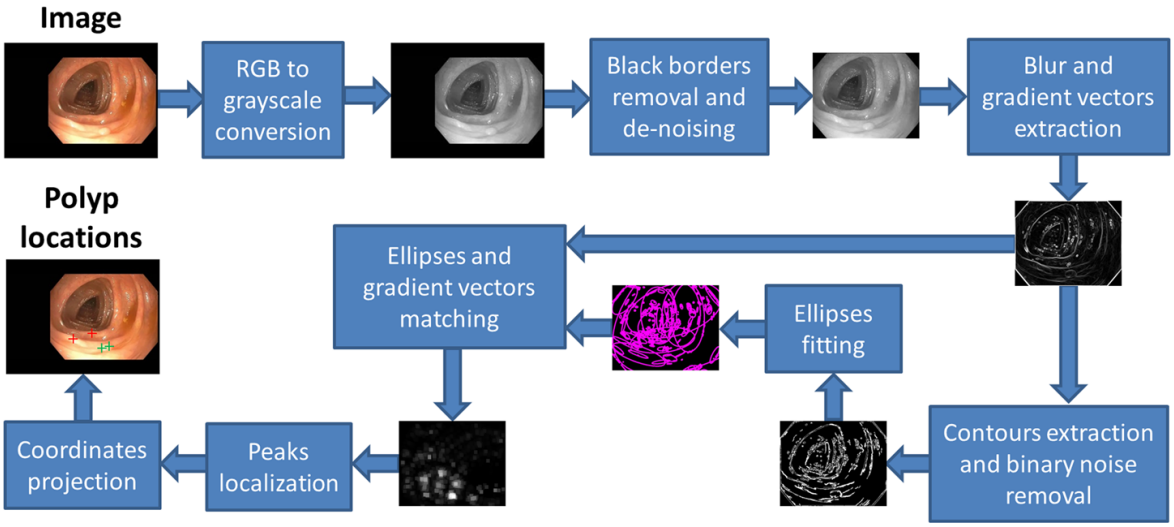


Fig. 4 Detailed steps of the basic EIR localization algorithm implementation

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.

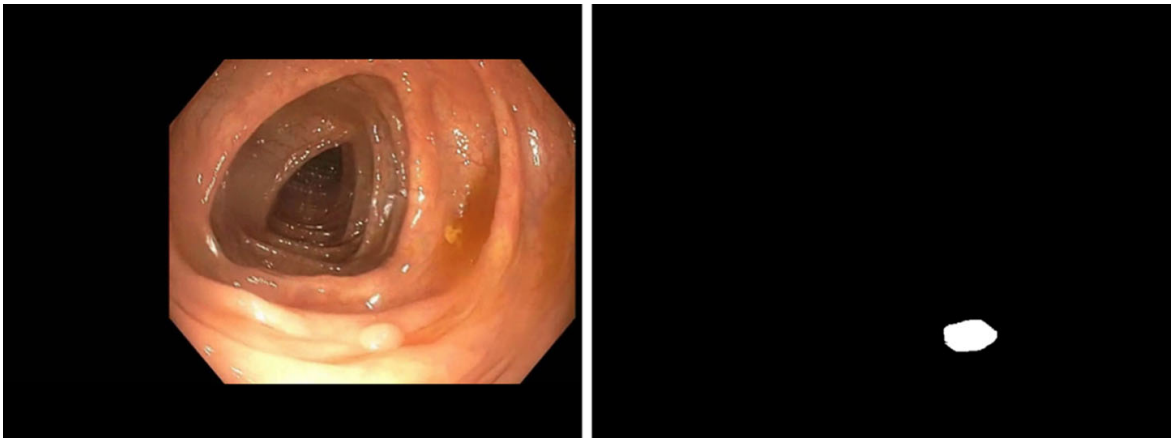
As initial models for both approaches, we used database models pre-trained on ImageNet [19] . Our custom training and testing data for the algorithms consists of frames and corresponding text files describing ground truth data with defined rectangular areas around objects: a JSON file for TensorBox and 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

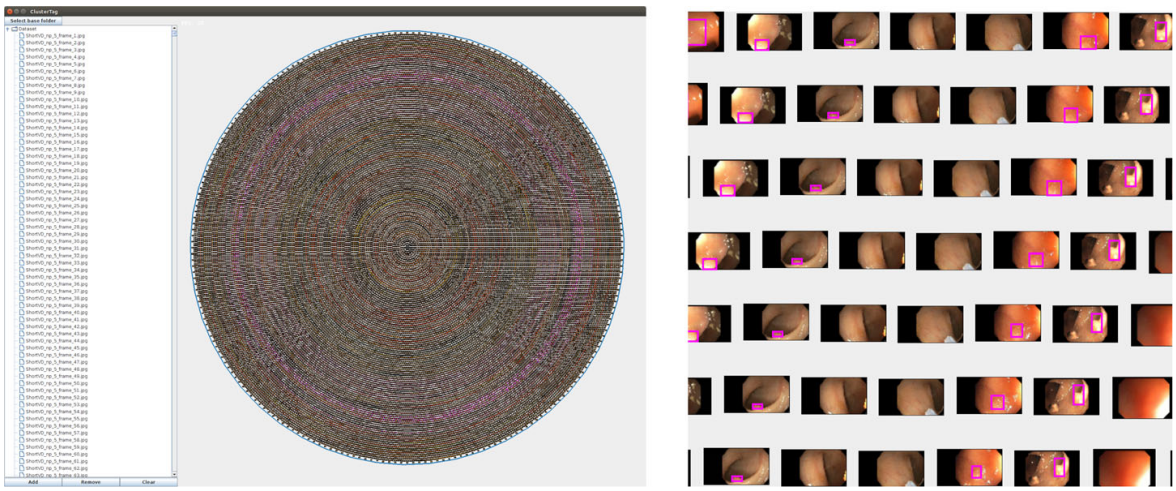


(a) Frame with polyp (b) Polyp ground truth

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

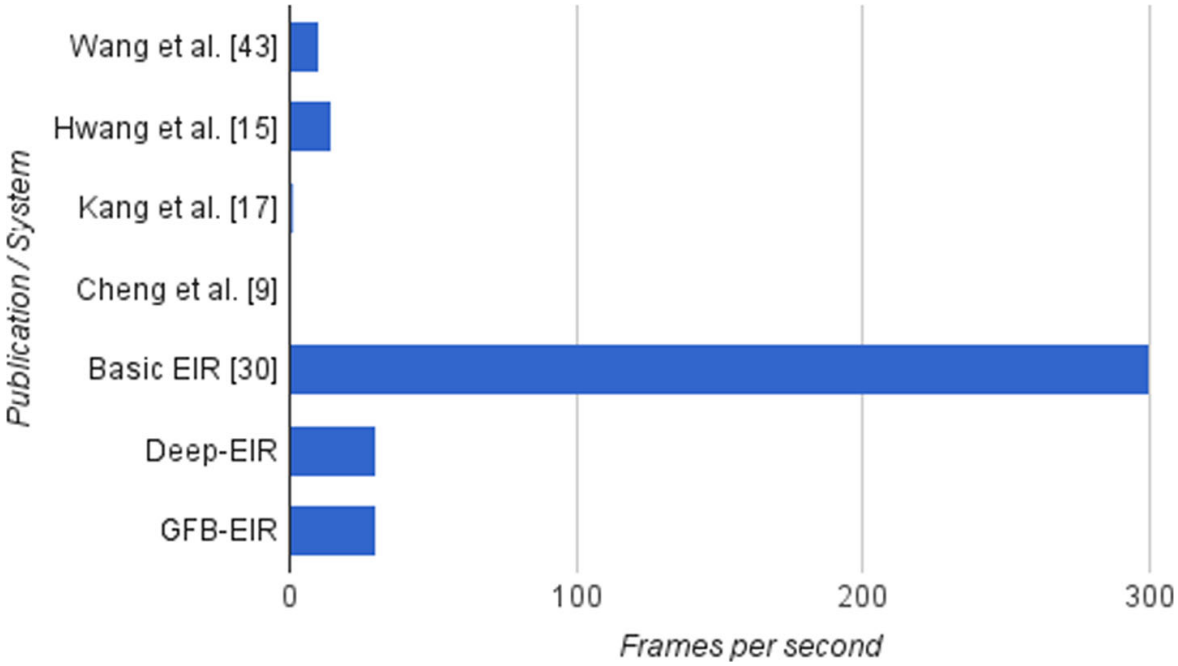


Fig. 7 The chart shows a comparison of different GI tract finding detection approaches. The presented Deep-EIR and multi-class global-feature-based EIR (GFB-EIR) systems show performance of 30 frames per second, which is higher comparing to other systems

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/

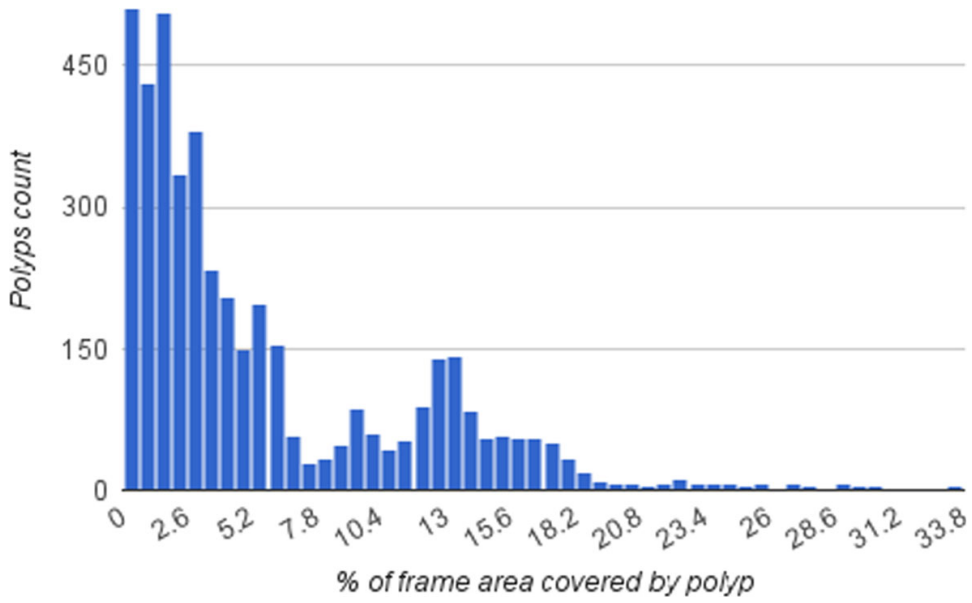


Fig. 8 The histogram shows huge variations in number of frame pixels, covered by polyp bodies, from very small up to one-third of full frame size across whole ASU-Mayo Clinic polyp database

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

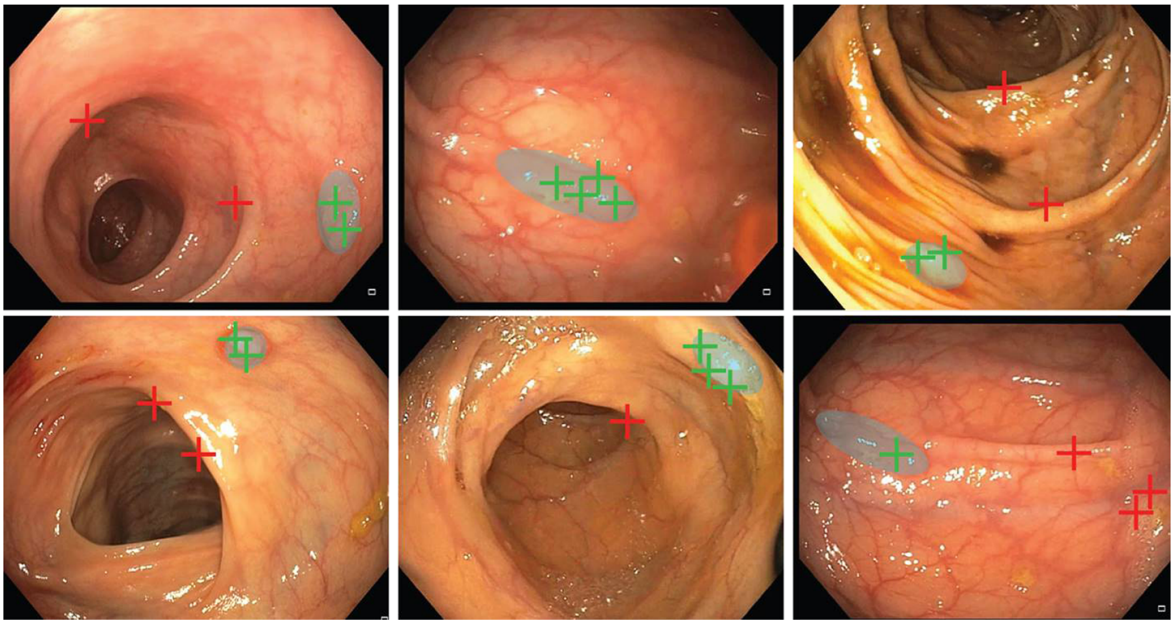


Fig. 9 Example of the polyp localization algorithm output. The current version of the algorithm produces four possible polyp locations per frame. The polyp location ground truth is marked by *light green* ellipses. The *green* crosses correspond to the true positives, the *red* correspond to the false 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

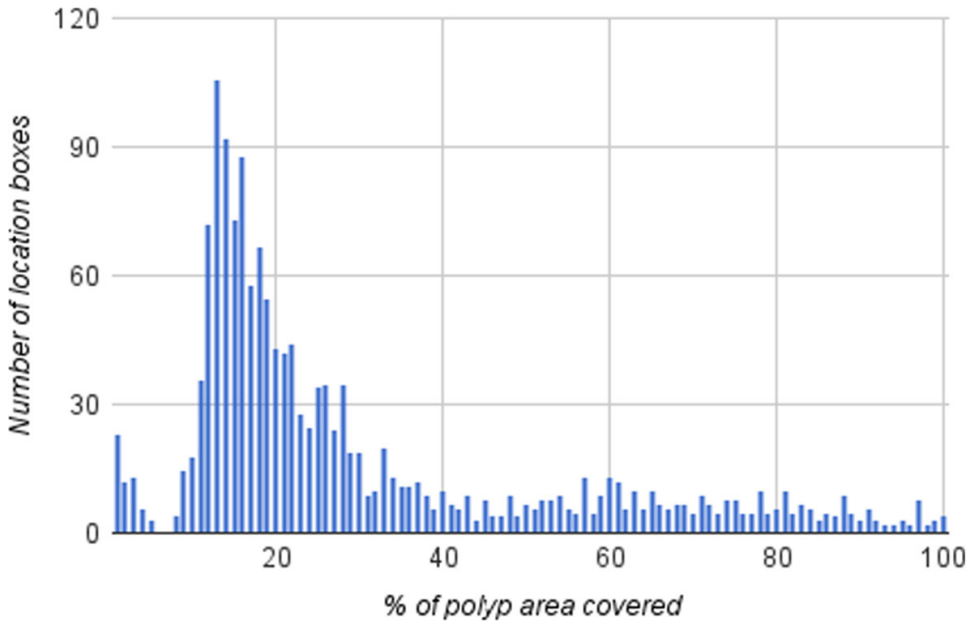


Fig. 10 The histogram shows polyps area coverage by location boxes found by the TensorBox localization algorithm with the maximum around 16% coverage rate with following exponential decrease to the almost constant level. The low number of found location rectangles around zero coverage rate is an evidence 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.

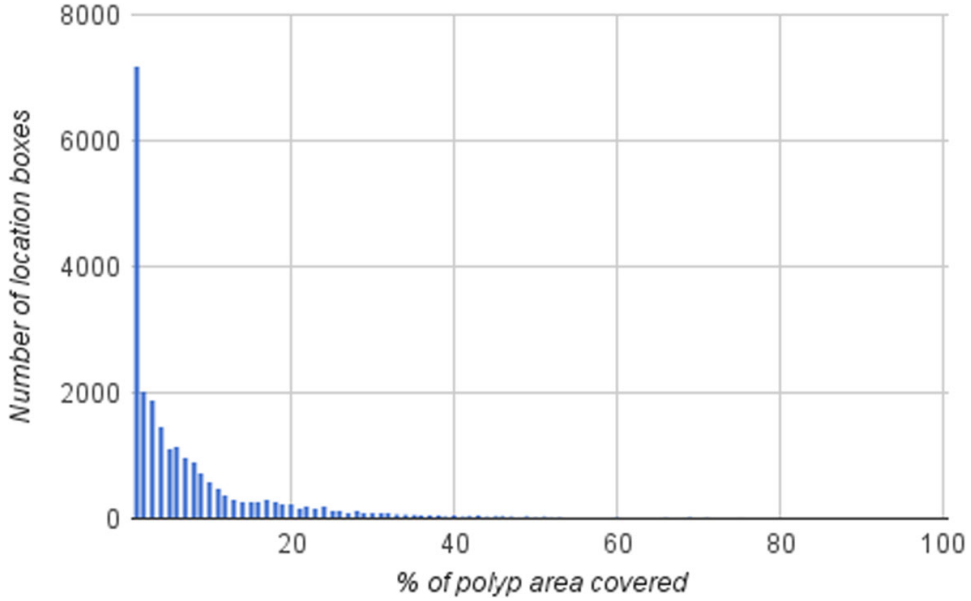


Fig. 11 The histogram shows polyps area coverage by location boxes found by the Darknet-YOLO localization algorithm with near to exponential distribution for coverage rate higher than 10%. The higher number of found location rectangles around zero coverage rate gives clear indications that algorithm output unfiltered results

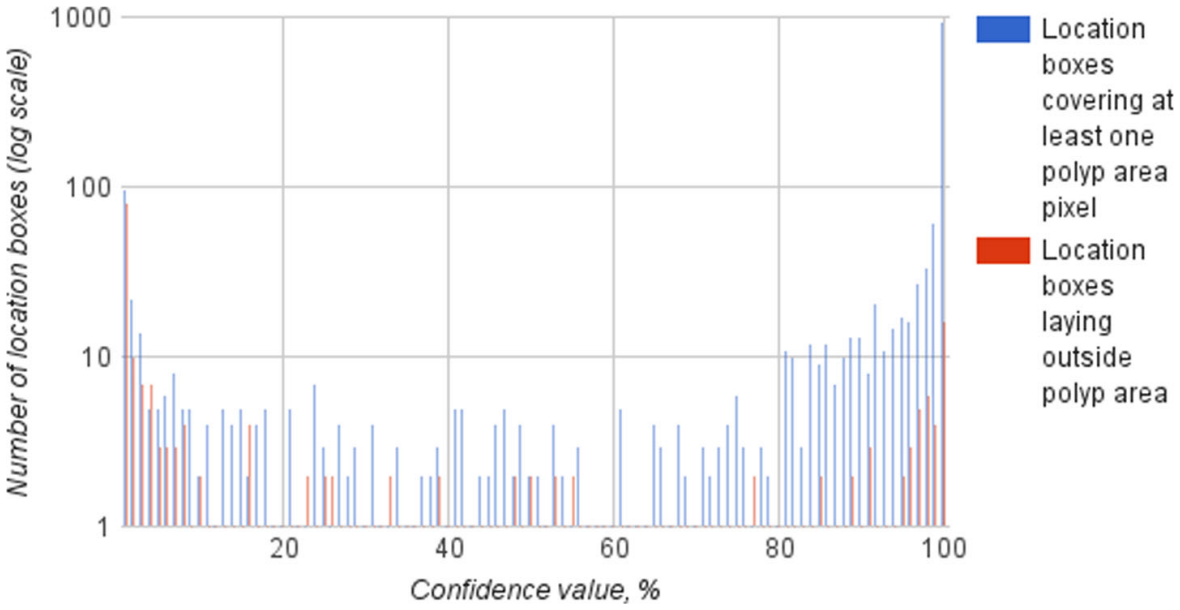


Fig. 12 The histogram shows confidence values for location boxes found by the TensorBox localization algorithm. It shows the relatively low number of found locations with high number of highly-confident locations

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

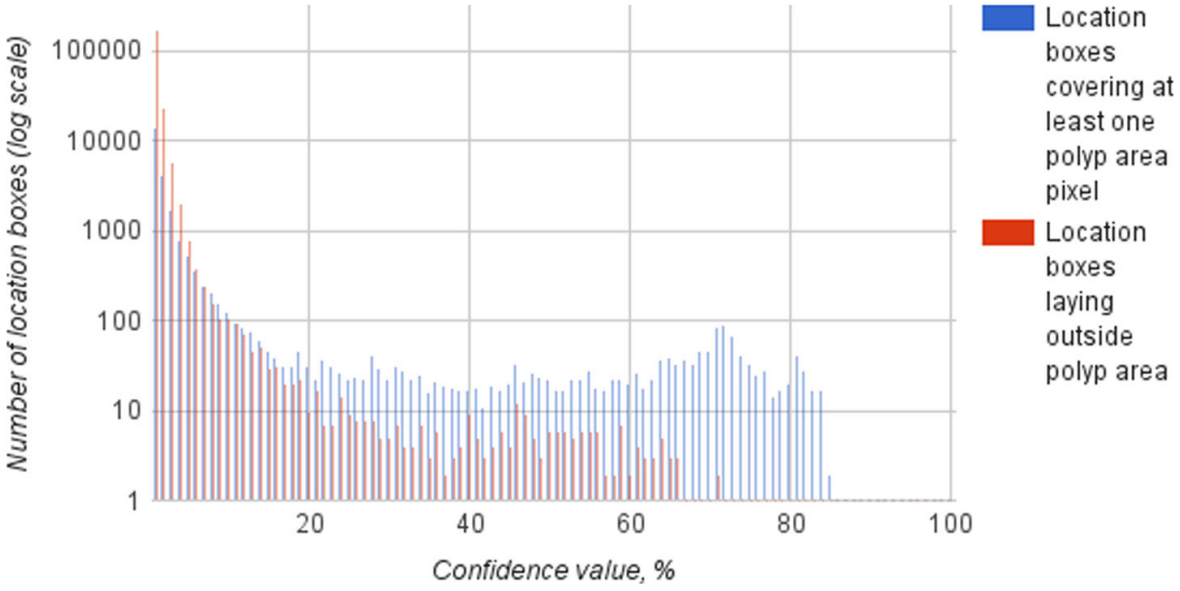


Fig. 13 The histogram shows confidence values and polyps area coverage by location boxes found by the Darknet-YOLO localization algorithm. It shows the expected high number of low-confident locations

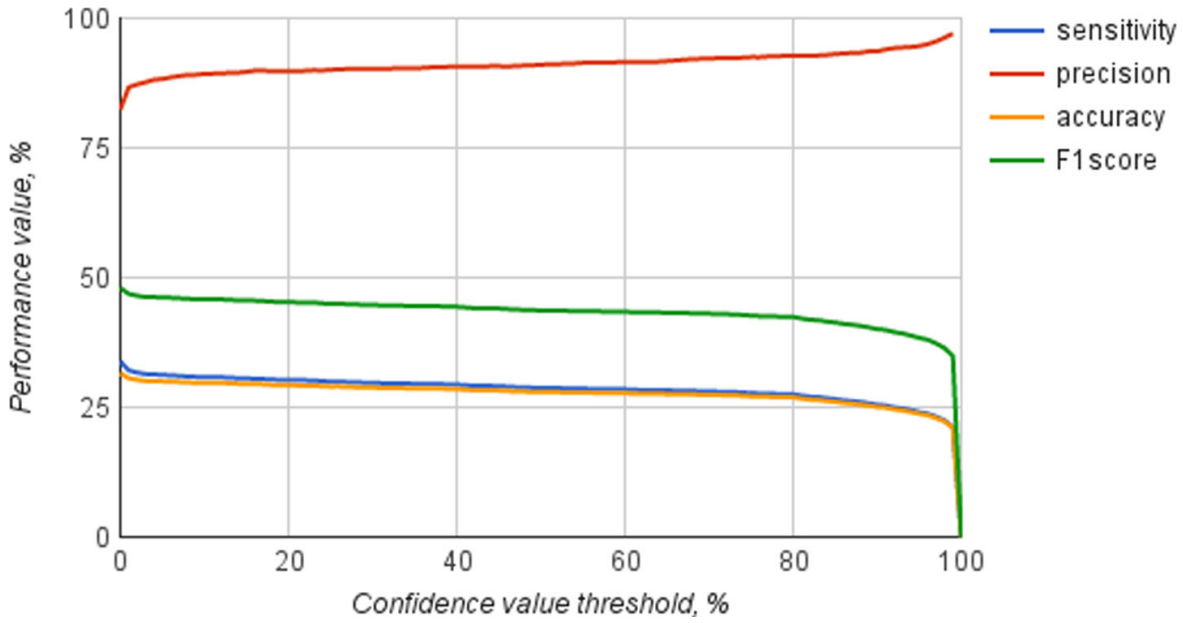


Fig. 14 The graphs show TensorBox localization algorithms performance for different confidence threshold values with no significant visible dependency. The maximum accuracy reaches level of 31.6% for zero-confidence threshold value with the polyp miss rate of 66.2%

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.

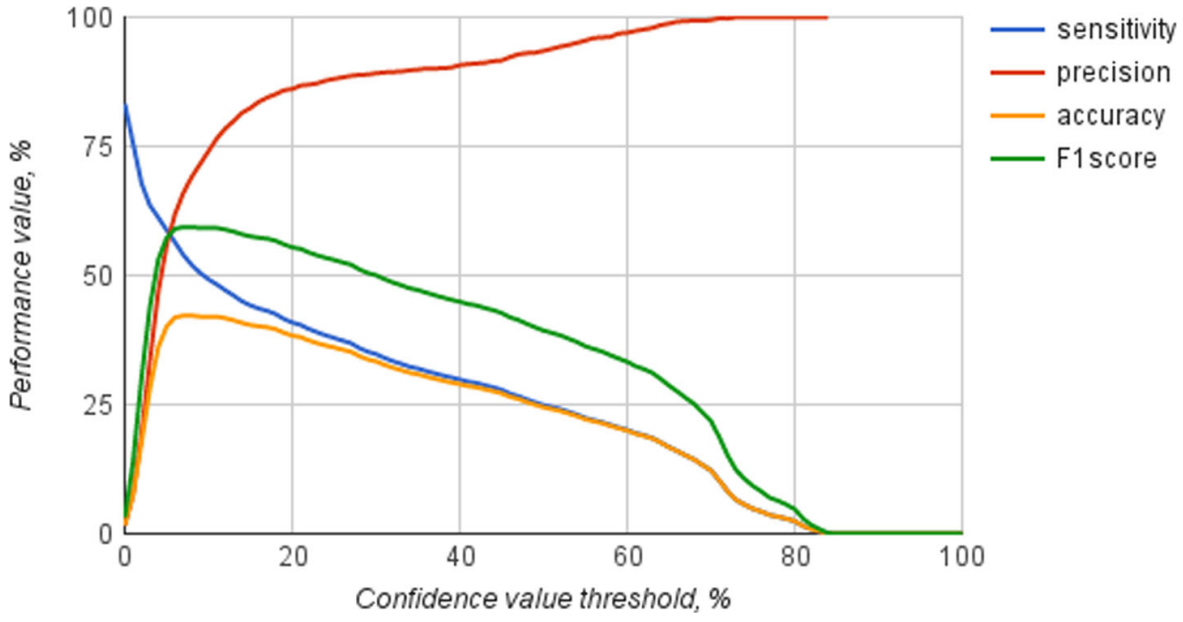


Fig. 15 The graphs show Darknet-YOLO localization algorithms performance for different confidence threshold values with good response to threshold value adjusting. The maximum accuracy reaches level of 42.2% for confidence threshold value of 8% with the polyp miss rate of 47.9%

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

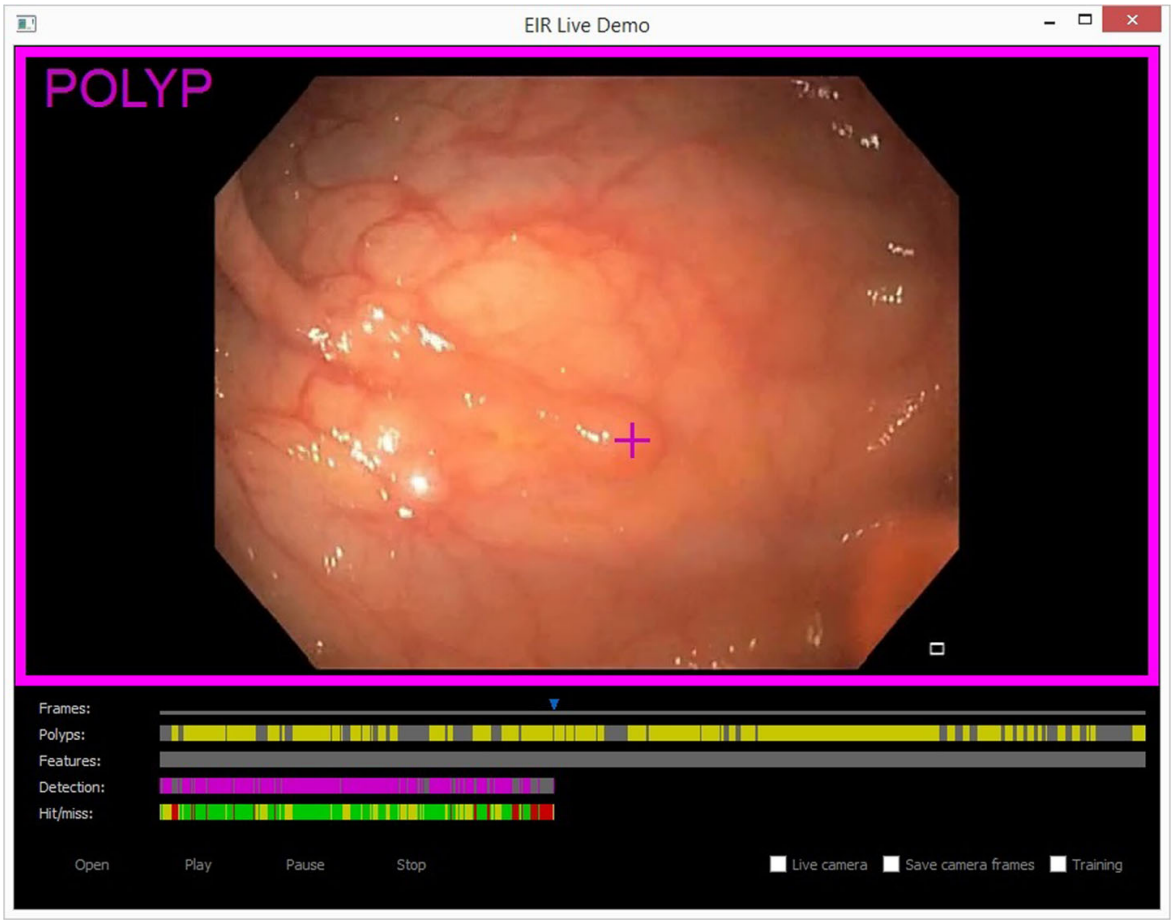


Fig. 16 A screenshot of the live system showing the combination of the visual information from the endoscope with feedback information from the detection and localization system. The pink fame surrounding background shows a positive detection. The name of the detected finding is shown in the frame auxiliary 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.

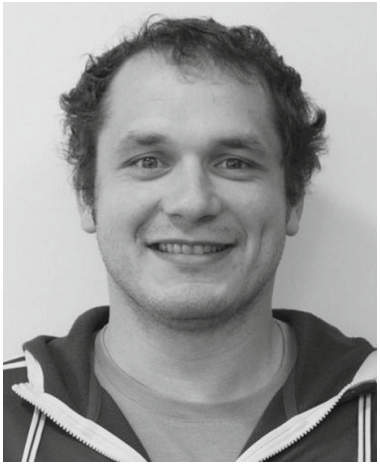
References

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

Paper XII

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

KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection

Konstantin Pogorelov
Simula Research Laboratory, Norway
University of Oslo, Norway

Kristin Ranheim Randel
Cancer Registry of Norway
University of Oslo, Norway

Carsten Griwodz
Simula Research Laboratory, Norway
University of Oslo, Norway

Sigrun Losada Eskeland
Bærum Hospital, Norway

Thomas de Lange
Bærum Hospital, Norway
Cancer Registry of Norway

Dag Johansen
UiT-The Arctic University of Norway

Concetto Spampinato
University of Catania, Italy

Duc-Tien Dang-Nguyen
Dublin City University, Ireland

Mathias Lux
University of Klagenfurt, Austria

Peter Thelin Schmidt
Karolinska Institutet, Solna, Sweden
Karolinska Hospital, Sweden

Michael Riegler
Simula Research Laboratory, Norway
University of Oslo, Norway

Pål Halvorsen
Simula Research Laboratory, Norway
University of Oslo, Norway

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.

ACM Reference format:
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, and Pål Halvorsen. 2017. KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of MMSys '17, Taipei, Taiwan, June 20–23, 2017,* 6 pages.
https://doi.org/http://dx.doi.org/10.1145/3083187.3083212

1 INTRODUCTION

The human digestive system may be affected by several diseases. As an example, three of the eight most common cancers overall are located in the gastrointestinal (GI) tract (figure 1). Altogether esophageal, stomach and colorectal cancer accounts for about 2.8 million new cases and 1.8 million deaths per year [40]. Endoscopic examinations (figures 2(a) and 2(b)) are the gold standards for investigation of the GI tract. Gastroscopy is an examination of the upper GI tract including esophagus, stomach and first part of small bowel, while colonoscopy covers the large bowel (colon) and rectum. Both these examinations are real-time video examinations of the inside of the GI tract by use of digital high definition endoscopes (figures 2(c)). Endoscopic examinations are resource demanding and requires both expensive technical equipment and trained personnel.

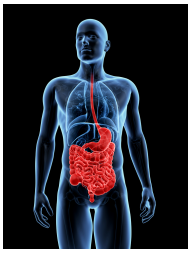


Figure 1: GI tract (shutterstock).

For colorectal cancer prevention, endoscopic detection and removal of possible precancerous lesions are essential. Adenoma detection is therefore considered to be an important quality indicator in colorectal cancer screening. However, the ability to detect adenomas varies between doctors, and this may eventually affect the individuals' risk of getting colorectal cancer [19].

Endoscopic assessment of severity and sub-classification of different findings may also vary from one doctor to another. Accurate grading of diseases are important since it may influence decision-making on treatment and follow-up [4, 11, 16]. For example, the degree of inflammation directly affects the choice of therapy in inflammatory bowel diseases (IBD) [37]. An objective and automated scoring system would therefore be highly welcomed.

Automatic detection, recognition and assessment of pathological findings will probably contribute to reduce inequalities, improve quality and optimize use of scarce medical resources. Furthermore, since endoscopic examinations are real-time investigations, both normal and abnormal findings have to be recorded and documented within written reports. Automatic report generation would probably contribute to reduce doctors' time required for paperwork and thereby increase time to patient care. Reliable and careful documentation with the use of minimal standard terminology (MST) [1]

This work is founded by the Norwegian FRINATEK project "EONS" (#231687).
Contact author: Konstantin Pogorelov, email: konstantin@simula.no .

MMSys '17, June 20–23, 2017, Taipei, Taiwan

ACM ISBN 978-1-4503-5002-0/17/06.
https://doi.org/http://dx.doi.org/10.1145/3083187.3083212

(a) Colonoscopy (b) Gastroscopy (c) A colonoscope

Figure 2: Various types of endoscopy examinations.

may also contribute to improved patient follow-up and treatment. To our knowledge, a standardized and automatic reporting system that ensure high quality endoscopy reports does not exist.

In order to make the health care system more scalable and cost effective, basic research in the intersection between computer science and medicine must go beyond traditional medical imaging by combining this area with multimedia data analysis and retrieval, artificial intelligence, and distributed processing. Next-generation medical big-data applications are a frontier for innovation, competition and productivity, where there are currently large initiatives both in the EU [26] and the US [24]. In the area of multimedia research, people are starting to see the synergies between multimedia and medical systems [31]. For automatic algorithmic detection of abnormalities in the GI tract, there have been many proposals from various research communities. For example, many systems present promising results for polyp detection [3, 5, 9, 18, 21, 23, 32, 38, 39, 41] reaching high precision and recall scores. However, the results are hard to reproduce due to lack of available medical data, i.e., the work listed above all use different data sets ranging from 35 to 1.8 million images/video frames.

In our earlier work [27, 32, 33], we have used the two only usable, publicly available GI tract datasets: the ASU-Mayo Clinic polyp database [35] and the CVC-ColonDB colonoscopy video database [7]. The ASU-Mayo dataset consists of training and test sets of images and videos with corresponding ground truth showing the exact polyp location areas. This is currently the biggest available dataset consisting of 20 videos from standard colonoscopies with a total of 18,781 frames and different resolution up to full HD. However, the images in this dataset are very similar raising the challenge of overfitting, and currently, the use of the dataset is restricted. The CVC-ColonDB dataset consists of images and videos partially covered by corresponding ground truth showing the exact polyp location areas. This is currently the second biggest available dataset consisting of 15 small videos from standard colonoscopies with a total of 1,200 frames and 300 frames with the region of interest marked. The resolution is $500x574$ pixels. Furthermore, both these datasets contain only one endoscopic finding (polyps). In this paper, we therefore publish Kvasir our multi-class dataset[1] from the Vestre Viken Health Trust (Norway) containing not only polyps, but also two other findings, two classes related to polyp removal and three anatomical landmarks in the GI tract.

2 DATA COLLECTION

The data is collected using equipment as shown in figure 2(c) at Vestre Viken Health Trust (VV) in Norway. The VV 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. Furthermore, the images are carefully annotated by one or more medical experts from VV and the Cancer Registry of Norway (CRN). The CRN provides new knowledge about cancer through research on cancer. It is part of South-Eastern Norway Regional Health Authority and is organized as an independent institution under Oslo University Hospital Trust. CRN is responsible for the national cancer screening programmes with the goal to prevent cancer death by discovering cancers or pre-cancerous lesions as early as possible.

[1] http://datasets.simula.no/kvasir

3 DATASET DETAILS

The initial Kvasir dataset consists of 4,000 images, annotated and verified by medical doctors (experienced endoscopists), including 8 classes showing anatomical landmarks, phatological findings or endoscopic procedures in the GI tract, i.e., 500 images for each class. The number of images is sufficient to be used for different tasks, e.g., image retrieval, machine learning, deep learning and transfer learning, etc. [2, 12, 28]. The anatomical landmarks are Z-line, pylorus and cecum, while the pathological finding includes esophagitis, polyps and ulcerative colitis. In addition, we provide two set of images related to removal of polyps, the "dyed and lifted polyp" and the "dyed resection margins". The dataset consist of the images with different resolution from $720x576$ up to $1920x1072$ pixels and organized in a way where they are sorted in separate folders named accordingly to the 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 (thus included), but must be handled with care for the detection of the endoscopic findings.

3.1 Anatomical Landmarks

An anatomical landmark is a recognizable feature within the GI tract that is easily visible through the endoscope. They are essential for navigating and as a reference point to describe the location of a given finding. The landmarks may 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 [30].

3.1.1 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. Recognition and assessment of the Z-line is important in order to determine whether disease is present or not. For example, this is the area where signs of gastro-esophageal reflux may appear. The Z-line is

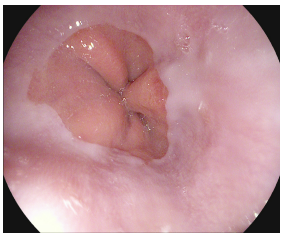


Figure 3: Z-line

also useful as a reference point when describing pathology in the esophagus.

3.1.2 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 4 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.

Figure 4: Pylorus

3.1.3 Cecum. The cecum the most proximal part of the large bowel. Reaching cecum is the proof for a complete colonoscopy and completion rate has shown to be a valid quality indicator for colonoscopy [6]. Therefore, recognition and documentation of the cecum is important. One of the characteristics hallmarks of 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 [29, 36]. Figure 5 shows an example of the appendiceal orifice visible as a crescent shaped slit, and the green picture in picture shows the scope configuration for cecal position.

Figure 5: Cecum

3.2 Phatological findings

A pathological finding in this context is an abnormal feature within the gastrointestinal tract. Endoscopically, it is visible as a damage or change in the normal mucosa. The finding may be signs of an ongoing disease or a precursor to for example cancer. Detection and classification of pathology is important in order to initiate correct treatment and/or follow-up of the patient.

3.2.1 Esophagitis. Esophagitis is an inflammation of the esophagus visible as a break in the esophageal mucosa in relation to the Z-line. Figure 6 shows an example with red mucosal tongues projecting up in the white esophageal lining. The grade of inflammation is defined by 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 treatment initiation to relieve symptoms and prevent further development of possible complications. Computer detection would be of special value in assessing severity and for automatic reporting.

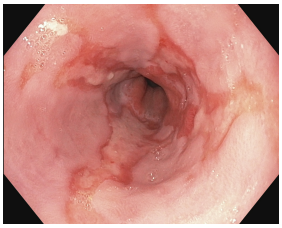


Figure 6: Esophagitis

3.2.2 Polyps. Polyps are lesions within the bowel detectable as mucosal outgrows. An example of a typical polyp is shown in figure 7. 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 development of colorectal cancer. Since polyps may be overlooked by the doctors, automatic detection would most likely improve examination quality. The green boxes within the image shows 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.

Figure 7: Polyp

3.2.3 Ulcerative colitis. Ulcerative colitis is a chronic inflammatory disease affecting the large bowel. The disease may have a large impact on quality of life, and diagnosis is mainly based on colonoscopic findings. The degree of inflammation varies from none, mild, moderate and 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 8 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.

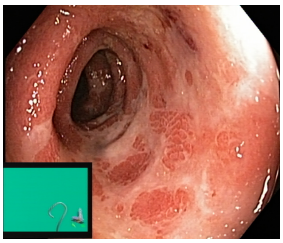


Figure 8: Ulcerative colitis

3.3 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 (EMR). 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. The 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 [17]. Computer detection of dyed polyps and the site of resection would be important in order to generate computer aided reporting systems for the future.

3.3.1 Dyed and Lifted Polyps. Figure 9 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 non-lifted areas that might indicate malignancy.

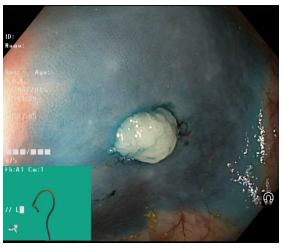


Figure 9: Dyed and Lifted Polyp

3.3.2 Dyed Resection Margins. The 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 worst case malignancy development. Figure 10 illustrates the resection site after removal of a polyp. Automatic recognition of the site of polyp removals are of value for automatic reporting systems and for computer aided assessment on completeness of the polyp removal.

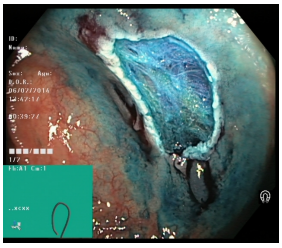


Figure 10: Dyed Resection Margin

4 APPLICATIONS OF THE DATASET

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 findings detection in the esophagus, stomach, bowel and rectum. Important results will include higher detection accuracies, reduced manual labor for medical personnel, reduced average cost, less patient discomfort and possibly increased willingness to undertake the examination. In the end, the improved screening will probably significantly reduce mortality and number of luminal GI disease incidents.

With respect to direct use in the multimedia research areas, 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 algoritmic analysis of images. Using the same collection of data, researchers can easier compare approaches and experimental results, and results can easier 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 developments of and experiments for 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.

In our work [27, 32, 33], we have for example conducted a leave-one-out cross-validation to evaluate our system. This is a method that assesses the generalization of a predictive model where the training and testing datasets are rotated, i.e., leaving out a single different non-overlapping item or portion for testing, and using the remaining items for training. This process is repeated until every item or portion has been used for testing exactly once [13]. Being one of the first medical multi-class datasets available to the multimedia community, we hereby invite and enable multimedia researcher into the medical domain of detection and retrieval.

5 SUGGESTED METRICS

Looking at the list of related work in this area, there are a lot of different metrics used, with potentially different names when used in the medical area and the computer science (information retrieval) area. Here, we provide a small list of the most important metrics. For future research, in addition to describing the dataset with respect to total number of images, total number of images in each class and total number of positives, it might be good to provide as many of the metrics below as possible in order to enable an indirect comparison with older work:

True positive (TP): The number of correctly identified samples. The number of frames with an endoscopic finding which correctly is identified as a frame with an endoscopic finding.

True negative (TN): The number of correctly identified negative samples, i.e., frames without an endoscopic finding which correctly is identified as a frame without an endoscopic finding.

False positive (FP): The number of wrongly identified samples, i.e., a commonly called a "false alarm". Frames without an endoscopic finding which is erroneously identified as a frame with an endoscopic finding.

False negative (FN): The number of wrongly identified negative samples. Frames without an endoscopic finding which erroneously is identified as a frame with an endoscopic finding.

Recall (REC): This metric is also frequently called *sensitivity*, *probability of detection* and *true positive rate*, and it is the ratio of samples that are correctly identified as positive among all existing positive samples:

$$recall = \frac{TP}{\#\ of\ all\ positives} = \frac{TP}{TP + FN}$$

Precision (PREC): This metric is also frequently called the *positive predictive value*, and shows the ratio of samples that are correctly identified as positive among the returned samples (the fraction of retrieved samples that are relevant):

$$precision = \frac{TP}{\#\ of\ all\ returned\ samples} = \frac{TP}{TP + FP}$$

Specificity (SPEC): This metric is frequently called the *true negative rate*, and shows the ratio of negatives that are correctly identified as such (e.g., the fraction of frames without an endoscopic finding are correctly identified as a negative result):

$$specificity = \frac{TN}{\#\ of\ all\ negatives} = \frac{TN}{FP + TN}$$

Accuracy (ACC): The percentage of correctly identified true and false samples:

$$accuracy = \frac{TP + TN}{\#\ of\ samples\ in\ total}$$

Matthews correlation coefficient (MCC): MCC takes into account true and false positives and negatives, and is a balanced measure even if the classes are of very different sizes:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

F1 score (F1): A measure of a test's accuracy by calculating the harmonic mean of the precision and recall:

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

In addition to the above metrics, system performance metrics processing speed and resource consumption are of interest. In our work, we have used the achieved frame-rate (FPS) as a metric as real-time feedback is important.

6 BASELINE PERFORMANCE

We have here performed an initial multi-class detection experiment on Kvasir as a baseline for future experiments. We have experimented using various configurations of three different main approaches, i.e., classification using global features (GF), deep learning convolutional neural networks (CNN) and transfer learning in deep learning (TFL).

For the GF approaches, we extracted several image features for classification using the latest version of the Lire open source software [22], i.e., the extracted features are JCD, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram and Pyramid Histrogram of Oriented Gradients. For the 2 GF run, we combined JCD and Tamura resulting in a feature vector of 187. For the 6 GF run, we combined all extracted features resulting in a feature vector of 1186. We decided for these combinations based on our previous findings and experiments in [32]. We performed a simple early fusion of the features, and all extracted features are included in the dataset in the *arff* file format for reuse and reproducibility. We used the Random Forrest (RF) and Logistic Model Tree (LMT) classifiers provided in the Weka machine learning library [15].

For all deep learning implementations, we used Keras [10] with Google Tensorflow [2] as backend. For the two CNN runs, we trained two different CNNs from scratch, i.e., one with three convolution layers and one with six. As activation function, we used the rectified linear unit (ReLU) [14] and for pooling maxpooling. In all layers, we also included a 0.5 dropout, and the final classification step was performed using two dense layers with first ReLU and then Sigmoid as activation functions. Both networks were trained for 200 epochs using the Adam optimizer [20].

The TFL run is based on transfer learning [8] by re-training and fine-tuning the pre-trained Inception v3 model [34]. For the re-training, we followed a similar approach as presented in [12]. Firstly, we locked all the basic convolutional layers of the network and only retrained the two top dense classification layers. The dense layers were retrained for 1,000 epochs using the RMSprop optimizer that allows an adaptive learning rate during the training process. After that, fine-tuning of a subset of the convolutional layers was performed. We decided to apply the fine-tuning on the two top convolutional layers of the re-trained model. For this training step, we used the SGD optimizer with a low learning rate (to achieve the best effect in terms of speed and accuracy) [25].

Table 1: Classification performance in terms of weighted average (2-folded) using the metrics described above.

Method	PREC	REC	SPEC	ACC	MCC	F1	FPS
6 Layer CNN	0.661	0.640	0.953	0.914	0.602	0.651	43
3 Layer CNN	0.589	0.408	0.890	0.959	0.430	0.453	45
Inception v3 TFL	0.698	0.689	0.957	0.924	0.649	0.693	66
2 GF Random Forrest	0.713	0.715	0.959	0.928	0.672	0.711	333
2 GF Logistic Model Tree	0.706	0.707	0.958	0.926	0.664	0.705	210
6 GF Random Forrest	0.732	0.732	0.962	0.933	0.692	0.727	105
6 GF Logistic Model Tree	0.748	0.748	0.964	0.937	0.711	0.747	80
Baseline (JCD Random Forrest)	0.708	0.710	0.958	0.927	0.666	0.706	370
Baseline (Random/Majority)	0.016	0.125	0.000	0.016	0.666	0.000	-

Table 2: Confusion matrix for both cross validated folds for the 6 GF LMT experiment in table 1. The classes are Esophagitis (A), Dyed and Lifted Polyps (B), Dyed Resection Margins (C), Cecum (D), Pylorus (E), Z-line (F), Polyps (G) and Ulcerative colitis (H). The test set in each fold contains 250 images for each class.

		Detected class							
		A	B	C	D	E	F	G	H
Actual class	A	**198/177**	0/0	0/0	0/0	3/8	49/64	0/1	0/0
	B	0/0	**139/149**	104/92	4/0	0/0	1/0	1/7	1/2
	C	0/0	90/100	**154/148**	2/0	0/0	1/0	2/1	1/1
	D	0/0	0/1	0/0	**214/223**	0/0	0/0	30/18	6/8
	E	5/3	0/0	0/0	0/0	**235/227**	2/8	5/12	3/0
	F	64/33	0/0	0/0	0/0	6/6	**180/210**	0/0	0/1
	G	0/0	0/0	4/1	24/26	10/2	2/2	**169/178**	41/41
	H	1/0	2/0	1/0	18/8	3/1	1/1	32/44	**192/196**

The exact configurations of the CNN and TFL approaches are included in the dataset. We did not perform any data augmentation, such as cropping, for any of the approaches for this work. For the experiments, we split the dataset randomly in two equally sized subsets (training and testing) containing 250 images per class each. We also performed two-folded cross-validation by switching the training and testing and calculated the average. As baselines, we provide one using the RF classifier with the JCD feature and one based on the random/majority class.

Table 1 gives and overview of the results, and table 2 contains the confusion matrix for the best performing approach (6 GF with LMT) for a more detailed insight into the performance. We can see that all approaches would outperform the random and majority class baseline, which is presented in the last row. Our own baseline in the second last row is only outperformed by three approaches. The best performing approach is a combination of six global features and the LMT classifier with an overall F1 score of 0.747 and 80 FPS. The 6 layer CNN outperforms the 3 layer CNN in terms of detection performance but not in terms of speed. The TFL approach outperforms the two other deep learning based approaches, which we expected since our CNN parameters are not optimized and we trained over a rather small number of epochs. Nevertheless, even if we use very basic methods, the here presented results can be a good starting point for other researchers and used as baselines to benchmark other methods applied to the dataset. In short, we see that multi-class detection is much more challenging than single detection, and that some findings are harder to detect than others, indicating that there are great potentials for improvements and innovations in future medical multimedia research.

7 CONCLUSION

To enable (reproducible) research in the intersection between multimedia and medicine, on analysis of images and videos of the human

GI tract in particular, we have presented the Kvasir dataset. The dataset has been collected during real endoscopy examinations and sorted and analyzed by medical experts. Initially, it contains 8 classes of images of important lesions and landmarks found in the GI tract, but it will be continuously updated. Medical datasets are hard to find, and such a dataset enables multi-disciplinary retrieval and detection research in order to improve health care systems all over the world.

REFERENCES

[1] Lars Aabakken, Alan N Barkun, Peter B Cotton, Evgeny Fedorov, Masayuki A Fujino, katerina Ivanova, Shin ei Kudo, Konstantin Kuznetzov, Thomas de Lange, Koji Matsuda, Olivier Moine, BjÃűrn Rembacken, Jean-Francois Rey, Joseph Romagnuolo, Thomas Rösch, Mandeep Sawhney, Kenshi Yao, and Jerome D Waye. 2014. Standardized endoscopic reporting. *J. of Gastroenterology and Hepatology* 29, 2 (2014), 234–240.

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[3] Luís A Alexandre, Joao Casteleiro, and Nuno Nobreinst. 2007. Polyp detection in endoscopic video using SVMs. In *Proc. of PKDD*. 358–365.

[4] Y. Amano, N. Ishimura, K. Furuta, K. Okita, M. Masaharu, T. Azumi, T. Ose, K. Koshino, S. Ishihara, K. Adachi, and Y. Kinoshita. 2006. Interobserver Agreement on Classifying Endoscopic Diagnoses of Nonerosive Esophagitis. *Endoscopy* 38, 10 (2006), 1032–1035.

[5] Stefan Ameling, Stephan Wirth, Dietrich Paulus, Gerard Lacey, and Fernando Vilarino. 2009. Texture-based polyp detection in colonoscopy. In *Bildverarbeitung für die Medizin*. Springer, 346–350.

[6] Nancy N. Baxter, Rinku Sutradhar, Shawn S. Forbes, Lawrence F. Paszat, Refik Saskin, and Linda Rabeneck. 2011. Analysis of administrative data finds endoscopist quality measures associated with postcolonoscopy colorectal cancer. *Gastroenterology* 140, 1 (2011), 65–72.

[7] Jorge Bernal, F. Javier Sanchez, and Fernando Vilarino. 2012. Towards Automatic Polyp Detection with a Polyp Appearance Model. *Pattern Recognition* 45, 9 (2012), 3166–3182.

[8] Souad Chaabouni, Jenny Benois-Pineau, and Chokri Ben Amar. 2016. Transfer learning with deep networks for saliency prediction in natural video. In *Proc. of ICIP*. 1604–1608.

[9] Da-Chuan Cheng, Wen-Chien Ting, Yung-Fu Chen, Qin Pu, and Xiaoyi Jiang. 2008. 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*. 62–72.

[10] François Chollet. 2015. Keras: Deep learning library for theano and tensorflow. (2015). https://keras.io/ Accessed: 2017-04-19.

[11] Thomas de Lange, Stig Larsen, and Lars Aabakken. 2004. Inter-observer agreement in the assessment of endoscopic findings in ulcerative colitis. *BMC Gastroenterology* 4, 9 (2004).

[12] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.. In *Proc. of ICML*, Vol. 32. 647–655.

[13] Bradley Efron and Robert Tibshirani. 1997. Improvements on Cross-Validation: The .632+ Bootstrap Method. *J. Amer. Statist. Assoc.* 92, 438 (1997), 548–560.

[14] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405, 6789 (2000), 947–951.

[15] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

[16] Lorenza A Herrero, Wouter L Curvers, Frederike G van Vilsteren, Herbert Wolfsen, Krish Ragunath, Louis-Michel W Song, Rosalie C Mallant-Hent, Arnoud van Oijen, Pieter Scholten, Erik J Schoon, Ed B Schenk, Bas L Weusten, and Jacques G Bergman. 2013. Validation of the Prague C&M classification of Barrett's esophagus in clinical practice. *Endoscopy* 45, 11 (2013), 876–882.

[17] Joo Ha Hwang, Vani Konda, Barham K Abu Dayyeh, Shailendra S Chauhan, Brintha K Enestvedt, Larissa L Fujii-Lau, Sri Komanduri, John T Maple, Faris M Murad, Rahul Pannala, Nirav C. Thosani, and Subhas Banerjee. 2015. Endoscopic mucosal resection. *Gastrointestinal Endoscopy* 82, 2 (2015), 215–226.

[18] Sae Hwang, JungHwan Oh, W. Tavanapong, J. Wong, and P.C. de Groen. 2007. Polyp Detection in Colonoscopy Video using Elliptical Shape Feature. In *Proc. of ICIP*. 465–468.

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

[20] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Baopu Li and M.Q.-H. Meng. 2012. Tumor Recognition in Wireless Capsule Endoscopy Images Using Textural Features and SVM-Based Feature Selection. *IEEE Trans. Information Technology in Biomedicine* 16, 3 (May 2012), 323–329.

[22] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. 2016. LIRE: open source visual information retrieval. In *Proc. of MMSys*. Article no. 30.

[23] A.V. Mamonov, I.N. Figueiredo, P.N. Figueiredo, and Y.-H.R. Tsai. 2014. Automated Polyp Detection in Colon Capsule Endoscopy. *IEEE Transactions on Medical Imaging* 33, 7 (July 2014), 1488–1502.

[24] McKinsey Global Institute. 2013. The big-data revolution in US health care: Accelerating value and innovation. (2013). https://goo.gl/SqS5Dl Accessed: 2017-04-19.

[25] Jiquan Ngiam, Adam Coates, Ahbik Lahiri, Bobby Prochnow, Quoc V Le, and Andrew Y Ng. 2011. On optimization methods for deep learning. In *Proc. of ICML*. 265–272.

[26] PMLiVE (Dominic Tyer). 2014. European Commission forms EUR2.5bn big data partnership. (2014). https://goo.gl/NeKb7H Accessed: 2017-04-19.

[27] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Peter T Schmidt, Carsten Griwodz, Dag Johansen, Sigrun L Eskeland, and Thomas de Lange. 2016. GPU-accelerated Real-time Gastrointestinal Diseases Detection. In *Proc. of CBMS*.

[28] Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2008. Transfer learning for image classification with sparse prototype representations. In *Proc. of CVPR*. 1–8.

[29] Douglas K Rex, Philip S Schoenfeld, Jonathan Cohen, Irving M Pike, Douglas G Adler, M Brian Fennerty, John G Lieb, Walter G Park, Maged K Rizk, Mandeep S Sawhney, Nicholas J Shaheen, Sachin Wani, and David S Weinberg. 2015. Quality indicators for colonoscopy. *American J. of Gastroenterology* 110, 1 (2015), 72–90.

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

[31] 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, and Pål Halvorsen. 2016. Multimedia and Medicine: Teammates for Better Disease Detection and Survival. In *Proc. of ACM MM*. 968–977.

[32] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Thomas de Lange, Carsten Griwodz, Peter Thelin Schmidt, Sigrun L. Eskeland, and Dag Johansen. 2016. EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal Endoscopies. In *Proc. of CBMI*.

[33] Michael Riegler, Konstantin Pogorelov, Jonas Markussen, Mathias Lux, Håkon Kvale Stensland, Thomas de Lange, Carsten Griwodz, Pål Halvorsen, Dag Johansen, Peter T Schmidt, and Sigrun L. Eskeland. 2016. Computer Aided Disease Detection System for Gastrointestinal Examinations. In *Proc. of MMSys*.

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567* (2015).

[35] Nima Tajbakhsh, Suryakanth Gurudu, and Jianming Liang. 2015. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE Transactions on Medical Imaging* 35, 2 (feb 2015), 630–644.

[36] R. Valori, J-F. Rey, W. S. Atkin, M. Bretthauer, C. Senore, G. Hoff, E. J. Kuipers, L. Altenhofen, R. Lambert, and G. Minoli. 2012. 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 (2012), SE88–SE105.

[37] A.J. Walsh, A. Ghosh, A.O. Brain, O. Buchel, D. Burger, S. Thomas, L. White G.S., Collins, S. Keshav, and S.P.L. Travis. 2014. Comparing disease activity indices in ulcerative colitis. *Journal of Crohn's and Colitis* 8, 4 (2014), 318–325.

[38] Yi Wang, Wallapak Tavanapong, Johnson Wong, JungHwan Oh, and Piet C de Groen. 2014. Part-Based Multiderivative Edge Cross-Sectional Profiles for Polyp Detection in Colonoscopy. *IEEE Journal of Biomedical and Health Informatics* 18, 4 (2014), 1379–1389.

[39] Yi Wang, Wallapak Tavanapong, Johnny Wong, Jung Hwan Oh, and Piet C de Groen. 2015. Polyp-Alert: Near Real-time Feedback during Colonoscopy. *Computer methods and programs in biomedicine* 3 (2015), 164–179.

[40] World Health Organization - International Agency for Research on Cancer. 2012. Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. (2012). https://goo.gl/IgZpVl Accessed: 2017-04-19.

[41] Mingda Zhou, Guanqun Bao, Yishuang Geng, B. Alkandari, and Xiaoxi Li. 2014. Polyp detection and radius measurement in small intestine using video capsule endoscopy. In *Proc. of BMEI*. 237–241.

Paper XIII

Nerthus: A Bowel Preparation Quality Video Dataset

NERTHUS: A Bowel Preparation Quality Video Dataset

Konstantin Pogorelov
Simula Research Laboratory, Norway
University of Oslo, Norway

Kristin Ranheim Randel
Cancer Registry of Norway
University of Oslo, Norway

Thomas de Lange
Bærum Hospital, Norway
Cancer Registry of Norway

Sigrun Losada Eskeland
Bærum Hospital, Norway

Carsten Griwodz
Simula Research Laboratory, Norway
University of Oslo, Norway

Dag Johansen
UiT-The Arctic University of Norway

Concetto Spampinato
University of Catania, Italy

Mario Taschwer
University of Klagenfurt, Austria

Mathias Lux
University of Klagenfurt, Austria

Peter Thelin Schmidt
Karolinska Institutet, Solna, Sweden
Karolinska Hospital, Sweden

Michael Riegler
Simula Research Laboratory, Norway
University of Oslo, Norway

Pål Halvorsen
Simula Research Laboratory, Norway
University of Oslo, Norway

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.

ACM Reference format:
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, and Pål Halvorsen. 2017. NERTHUS: A Bowel Preparation Quality Video Dataset. In *Proceedings of MMSys '17, Taipei, Taiwan, June 20–23, 2017,* 5 pages.
https://doi.org/http://dx.doi.org/10.1145/3083187.3083216

This work is founded by the Norwegian FRINATEK project "EONS" (#231687).
Contact author: Konstantin Pogorelov, email: konstantin@simula.no .

MMSys '17, June 20–23, 2017, Taipei, Taiwan

ACM ISBN 978-1-4503-5002-0/17/06.
https://doi.org/http://dx.doi.org/10.1145/3083187.3083216

1 INTRODUCTION

The large bowel (figure 1), also named colon and large intestine, is the lower part of the human gastrointestinal (GI) tract. It may be affected by severe diseases including cancer and chronic inflammations. Bowel cancer (colorectal cancer) is currently the third most common cancer worldwide, accounting for nearly 1.4 million new cases and 700 000 cancer deaths in 2012 [9, 27]. The current gold standard for diagnostic and screening investigations of the large bowel is *colonoscopy*. This is a real-time video examination of the inside of the large bowel by use of a digital high definition endoscope. Such endoscopic examinations are resource demanding and require both expensive technical equipment and trained personnel.

Figure 1: The large bowel (image: shutterstock.com).

Furthermore, the efficiency of colonoscopy depends on sufficient bowel cleansing to visualize the gastric mucosa (a membrane that lines the GI tract), achieved by use of oral laxatives (substances that loosen stools and increase bowel movements) administrated prior to the procedure. The quality of bowel preparation has shown to influence both the colonoscopy completion rate and detection of possible precursors of cancer (e.g., adenomas, which are the benign tumor of epithelial tissue) [10, 26]. Adenoma detection rate (ADR), that is inversely associated with a patient's risk of developing colorectal cancer, has been proven to be dependent on quality of bowel preparation [7, 13]. Therefore, the degree of bowel preparation is considered to be a reliable quality measure for colonoscopy [16].

Quality of bowel preparation may also influence decisions on screening and follow-up intervals, since low-quality bowel preparation requires repeated colonoscopy [6]. An accurate description of the bowel cleanliness is therefore needed. 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 may contribute to reduce such inequalities and optimize use of medical resources. Since endoscopic examinations are real-time investigations, both normal and abnormal findings have to be recorded and documented within written reports. Thus, automatic report generation will probably contribute to reduce doctors' time required for paperwork and thereby increase time to patient care. To our knowledge, a standardized and automatic reporting system that ensures high quality endoscopy reports does not exist. Assessment of bowel cleanliness would be a valuable feature for such automatic endoscopy reporting in the future.

In this area of research, people start to see the synergies between multimedia and medical systems [22]. The development of real-time classification systems is a perfect match in the intersection of medicine, multimedia systems and image/video retrieval. Prototypes like the EIR [23] system targeting analysis of medical videos for detection of abnormalities may be an initial starting point. However, such systems require a lot of data for development, training and testing. To the best of our knowledge, no medical dataset exists for this type of data. In this paper, we therefore present the Nerthus dataset[1]. It contains 21 videos with a total number of 5,525 frames annotated and verified by medical doctors (experienced endoscopists). The videos are divided into four classes of predefined bowel-preparation qualities.

Our initial experiments indicate potential for improvement. In many cases, we are able to detect the annotated bowel cleansing quality. However, to deliver an automatic and reliable system for the endoscopy units, more work is needed. By providing the Nerthus dataset, we invite researchers to contribute in order to improve important systems for automatic assessment and reporting for GI endoscopy.

2 BOWEL PREPARATION QUALITY

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. In addition, such a traditional categorization relies on a global assessment of bowel cleanliness, which does not account for differences in cleansing between bowel segments. Poor quality of preparation in one segment may then result in low overall grading, despite an otherwise perfectly cleaned bowel. To minimize the inter-endoscopist variation, new score-based methods of assessing bowel cleanliness have been introduced during the last decade.

State-of-the-art scoring systems include the BBPS [3, 15] and the Ottawa bowel preparation scale (OBPS) [24]. Both these scales divide the bowel into three sections (right, middle and left) and score the bowel cleansing within each section according to a defined numeric scale. OBPS uses segmental scores ranges from 0 to 4 in addition to a global three-score fluid-quantity rating, which requires estimation of residual liquid. In contrast, the Boston bowel preparation scale (BBPS) [3, 15] uses only a four-point scoring system (ranges from 0 to 3). Figure 2 illustrates the segmental division of the large bowel used for bowel preparation assessment according to BBPS and OBPS.

[1] http://datasets.simula.no/nerthus

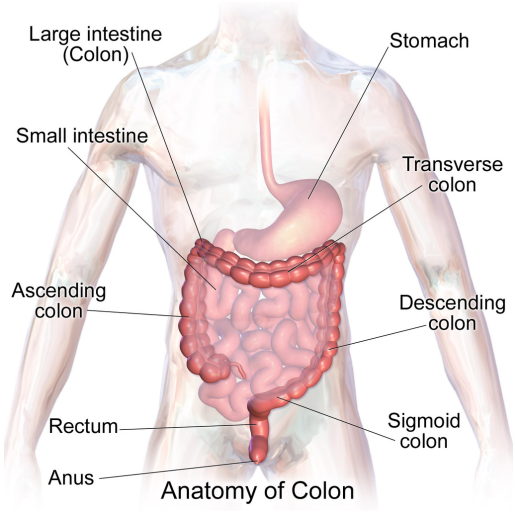


Figure 2: The segmental division of the large bowel: the right side includes the cecum and ascending colon, the transverse section of the colon includes the hepatic and splenic flexures, and the left side includes the descending colon, sigmoid colon, and rectum (image: WikiJournal of Medicine [2]).

In this paper, we use BBPS as this is probably best validated and most frequently used scoring system in both routine clinic and screening settings today [19]. The BBPS scale is tested and validated to assess the cleanliness at withdrawal. It does not take into account whether the endoscopist has performed any additional cleansing maneuvers, which reflects the actual practice of colonoscopy. The definition of the BBPS segmental scores are described in table 1. The segmental scores ranges from 0 to 3, where 0 is worse and 3 is the best quality of the bowel preparation [3, 15]. Examples for the different categories are shown in figure 4.

In real colonoscopy examinations, a segmental score is applied [15] to each of the three bowel segments and summed in a total score ranging from 0 to 9. In the Nerthus dataset, however, all videos

Table 1: The score points used in the Boston bowel preparation scale (BBPS) [3, 15] to define the degrees of bowel cleanliness.

Score	Description
0	Unprepared colon segment with mucosa not seen because of solid stool that cannot be cleared
1	Portion of mucosa of the colon segment seen, but other areas of the colon segment are not well seen because of staining, residual stool, and/or opaque liquid
2	Minor amount of residual staining, small fragments of stool, and/or opaque liquid, but mucosa of colon segment is seen well
3	Entire mucosa of colon segment seen well, with no residual staining, small fragments of stool, or opaque liquid

are recorded in the left part of the bowel. Automatic detection of scope position or total score calculation is thereby irrelevant for this dataset and only quality of bowel preparation by segmental scores are of value here. For the future development of automated systems, detecting position and assessment of total BBPS score will be of interest.

3 DATA COLLECTION

The data is collected using equipment as shown in figure 3 at Bærum hospital, Vestre Viken Hospital Trust in Norway. Furthermore, the videos are annotated by one or more medical experts from the Cancer Registry of Norway. A selection of the videos will in addition be annotated by several medical experts from Norway, Sweden, UK, US and Canada through a web based test. These video clips will be marked as the gold standard within dataset and will be released as an addition to the Nerthus dataset with higher quality regarding bowel preparation assessment.

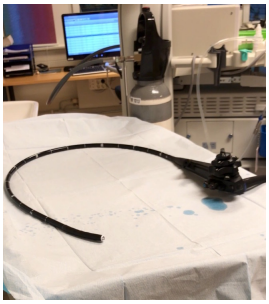


Figure 3: A colonoscope.

4 DATASET DETAILS

The Nerthus dataset consists of 21 videos with a total number of 5,525 frames, annotated and verified by medical doctors (experienced endoscopists), including 4 classes showing four-score BBPS-defined bowel-preparation quality (see figure 4). 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. [1, 8, 21]. The dataset consists of videos with resolution $720x576$ and is organized by sorting the videos into separate folders named according to their BBPS-bowel preparation quality score. 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 by use of 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, but must be handled with care for the bowel preparation quality assessment.

5 SUGGESTED METRICS

When reviewing related work in medical and the computer science areas, there are a lot of different metrics used, with potentially different names when used in the medical area and the computer science (information retrieval) area. For future research, in addition to describing the dataset with respect to total number of images, total number of images in each class and total number of positives, etc., it might be good to provide as many of the *recall* (also known as sensitivity and probability of detection, REC), *precision* (also known as the positive predictive value, PREC), *specificity* (also known as the true negative rate, SPEC), *accuracy* (ACC), *Matthews correlation coefficient* (MCC) and the *F1 score* (F1) metrics [20] as possible in order to enable a comparison with other work. In addition to the above metrics, processing speed and resource consumption are of interest. In our work, we have used the achieved frame-rate (FPS) as a metric as real-time feedback is important.

6 BASELINE PERFORMANCE

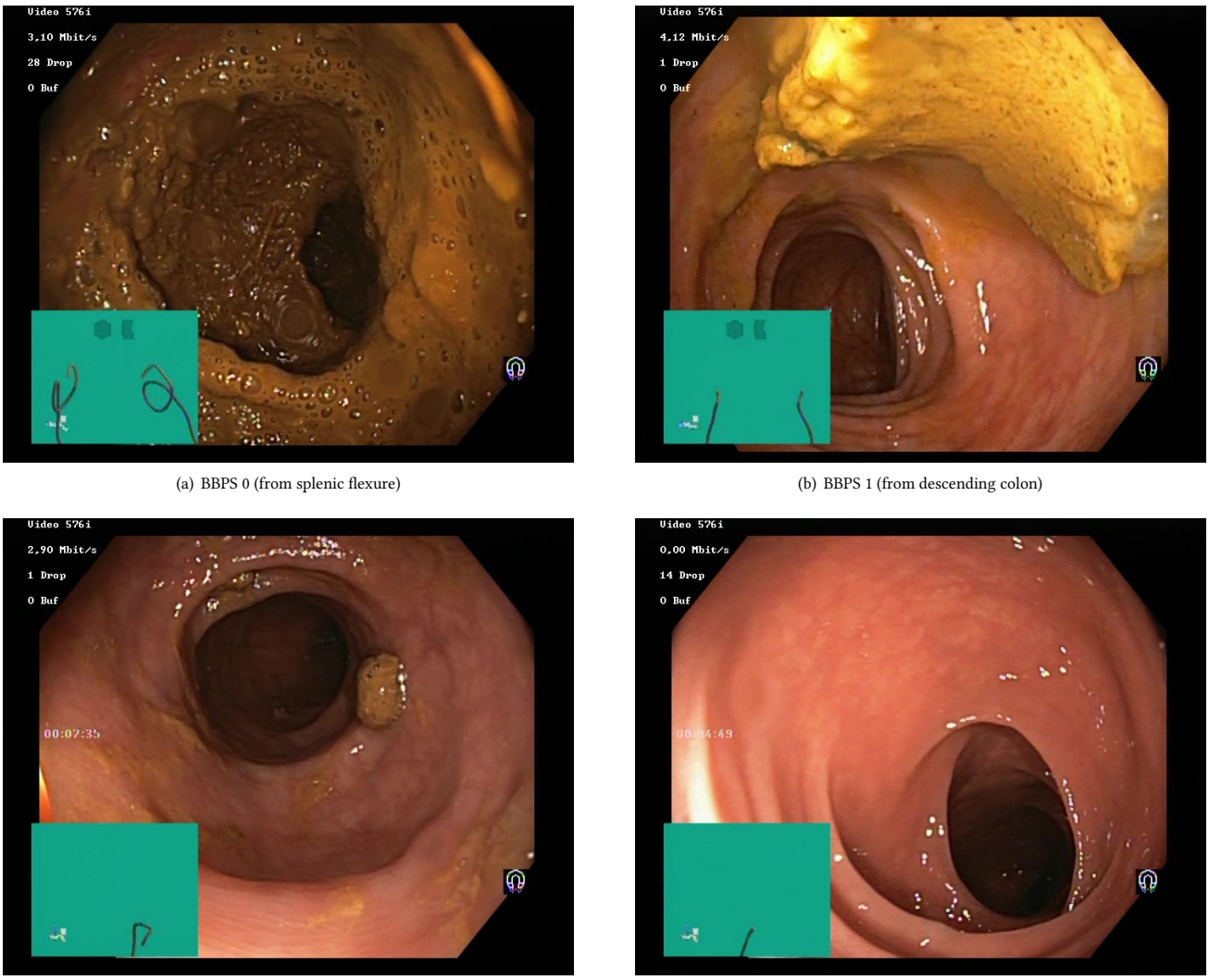
We have performed an initial multi-class detection experiment on Nerthus as a baseline for future experiments. We have experimented using various configurations of three different main approaches: classification using global features (GF), deep learning convolutional neural networks (CNN) and transfer learning in deep learning (TFL).

For the GF approaches, we extracted several image features for classification using the latest version of the Lire open source software [17]. The extracted features are JCD, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram and Pyramid Histrogram of Oriented Gradients. For a first GF run (2 GF), we combined JCD and Tamura, resulting in a feature vector of 187. For a second GF run (6 GF), we combined all extracted features, resulting in a feature vector of 1186 values. The decision of these combinations was based on our previous findings and experiments in [23]. Subsequently, we performed a simple early fusion of the features. All extracted features are included in the dataset in the *arff* file format for reuse and reproducibility. We used the Random Forrest (RF) and Logistic Model Tree (LMT) classifiers provided in the Weka machine learning library [12].

For all deep learning implementations, we used Keras [5] with Google Tensorflow [1] as backend. For the two CNN runs, we trained two different CNNs from scratch, i.e., one with three convolution layers and one with six. As activation function, we used the rectified linear unit (ReLU) [11], and for pooling, we used max-pooling. In all layers, we also included a 0.5 dropout, and the final classification step was performed using two dense layers with first ReLU and then Sigmoid as activation functions. The networks were trained for 200 epochs using the Adam optimizer [14].

The TFL run is based on transfer learning [4] by re-training and fine-tuning the pre-trained Inception v3 model [25]. For the re-training, we followed a similar approach to the one presented in [8]. Firstly, we locked all the basic convolutional layers of the network and only retrained the two top dense classification layers. The dense layers were retrained for 100 epochs using the RMSprop optimizer that allows an adaptive learning rate during the training process. This is was done for 100 epochs. After that, fine-tuning of a subset of the convolutional layers was performed. We decided to apply the fine-tuning on the two top convolutional layers of the re-trained model. For this training step, we used the SGD optimizer with a low learning rate (to achieve the best effect in terms of speed and accuracy) [18].

The exact configurations of the CNN and TFL approaches are included in the dataset. We did not perform any data augmentation, such as cropping, for any of the approaches for this work. For the experiments, we first split all the dataset videos into non-overlapping five-second-long video segments resulting in a new set of videos consisting of 52 files. Next, we split the new set randomly into two training and testing subsets, sized as equally as possible depending



(a) BBPS 0 (from splenic flexure)

(b) BBPS 1 (from descending colon)

(c) BBPS 2 (from sigmoid colon)

(d) BBPS 3 (from rectum)

Figure 4: Sample images for each bowel preparation ("cleanliness") score according to BBPS.

on the number of videos available for each class. The subsets contain 27 and 25 videos each. Then, we extracted all the frames from the videos, which gave us two new subsets containing 2800 and 2725 frames each, and applied our algorithms to the new frame sets. We also performed two-folded cross-validation by switching the training and testing sets and calculated the average. As baselines, we provide results of two runs: the RF GF-based classifier with the JCD feature and the random/majority-based classifier.

Table 2 gives an overview of the results, and table 3 contains the confusion matrix for the best performing approach (6 GF with LMT) for a more detailed insight into the performance. We observe that all approaches would outperform the random and majority class baseline, which is presented in the last row. Our own baseline in the second last row is only outperformed by two 6 GF-based approaches and the 6 layer CNN approach. The best performing approach is a combination of six global features and the LMT classifier with an overall F1 score of 0.899 and 77 FPS. The 6 GF RF approach is slightly faster, but less accurate, than 6 GF LMT. The 2 GF approaches

Table 2: Classification performance in terms of weighted average (2-folded) using the metrics described above.

Method	PREC	REC	SPEC	ACC	MCC	F1	FPS
6 Layer CNN	0.856	0.852	0.952	0.854	0.805	0.854	42
3 Layer CNN	0.811	0.694	0.937	0.772	0.621	0.742	40
Inception v3 TFL	0.751	0.745	0.918	0.748	0.665	0.748	61
2 GF Random Forrest	0.792	0.774	0.849	0.847	0.647	0.769	310
2 GF Logistic Model Tree	0.744	0.737	0.862	0.825	0.594	0.737	200
6 GF Random Forrest	0.885	0.866	0.895	0.913	0.801	0.860	101
6 GF Logistic Model Tree	0.901	0.901	0.960	0.949	0.863	0.899	77
Baseline (JCD Random Forrest)	0.805	0.794	0.870	0.861	0.679	0.79	330
Baseline (Random/Majority)	0.240	0.489	0.512	0.652	0.000	0.322	-

Table 3: Confusion matrix for both cross validated folds for the 6 GF LMT experiment in table 2. The classes correspond to bowel-preparation quality in BBPS score points from zero to three.

		Detected class				*Number of images*
		0	1	2	3	*in the test set*
Actual class	0	**248/236**	1/0	1/14	0/0	250/250
	1	14/9	**1186/1457**	23/2	2/7	1225/1475
	2	1/0	4/8	**398/423**	97/44	500/475
	3	0/0	104/16	132/68	**514/516**	750/600

gives almost the same detection, while 2 GF RF-based approach is faster and performs slightly better. The 6 layer CNN approach performs almost as good as 6 GF RF-based approach, while it is more than 2 times slower. The 3 layers CNN and TFL approaches performs almost equally in terms of detection performance, while the TFL output is slightly better balanced in terms of precision and recall. The performance of the 2 GF approaches and 3 layers CNN and TFL approaches is almost the same. The relatively low performance of the deep learning based approaches is expected since neural network parameters are not optimized and we trained over a rather small number of epochs. Nevertheless, even if we use very basic methods, our results can be a good starting point for other researchers and used as baselines to benchmark other methods applied to the dataset.

7 CONCLUSION

Adequate bowel preparation (cleansing) is required to achieve high quality colonoscopy examinations. Despite the use of reliable and validated bowel preparation assessment scales, the grading may vary from one doctor to another. By providing the Nerthus dataset, we invite multimedia researchers to contribute in the medical field by making systems that automatically and consistently can evaluate the quality of bowel cleansing. Innovations in this area contributing with computer-aided assessment and automatic reporting may potentially improve the medical field of GI endoscopy.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Blausen.com staff. 2014. Medical gallery of Blausen Medical 2014. *WikiJournal of Medicine* 1, 2 (2014).

[3] AH Calderwood and BC Jacobson. 2010. Comprehensive validation of the Boston Bowel Preparation Scale. *Gastrointestinal endoscopy* 72, 4 (2010), 686–692.

[4] Souad Chaabouni, Jenny Benois-Pineau, and Chokri Ben Amar. 2016. Transfer learning with deep networks for saliency prediction in natural video. In *Proc. of ICIP*. 1604–1608.

[5] François Chollet. 2015. Keras: Deep learning library for theano and tensorflow. (2015). https://keras.io/ Accessed: 2017-04-19.

[6] Brian T. Clark, Tarun Rustagi, and Loren Laine. 2014. What level of bowel prep quality requires early repeat colonoscopy: systematic review and meta-analysis of the impact of preparation quality on adenoma detection rate. *Gastroenterol Report* 109, 11 (2014), 1714–23.

[7] Douglas A. Corley, Christopher D. Jensen, Amy R. Marks, Wei K. Zhao, Jeffrey K. Lee, Chyke A. Doubeni, Ann G. Zauber, Jolanda de Boer, Bruce H. Fireman, Joanne E. Schottinger, Virginia P. Quinn, Nirupa R. Ghai, Theodore R. Levin, and Charles P. Quesenberry. 2014. Adenoma detection rate and risk of colorectal cancer and death. *New england journal of medicine* 370, 14 (2014), 1298–306.

[8] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.. In *Proc. of ICML*, Vol. 32. 647–655.

[9] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer* 136, 5 (2015), E359–86.

[10] Florian Froehlich, Vincent Wietlisbach, Jean-Jacques Gonvers, Bernard Burnand, and John-Paul Vader. 2005. Impact of colonic cleansing on quality and diagnostic yield of colonoscopy: the European Panel of Appropriateness of Gastrointestinal Endoscopy European multicenter study. *Gastrointestinal Endoscopy* 61, 3 (2005), 378–384.

[11] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405, 6789 (2000), 947–951.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

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

[14] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Edwin J Lai, Audrey H Calderwood, Gheorghe Doros, Oren K Fix, and Brian C Jacobson. 2009. The Boston bowel preparation scale: a valid and reliable instrument for colonoscopy-oriented research. *Gastrointestinal endoscopy* 69, 3 (2009), 620–625.

[16] Benjamin Lebwohl, Fay Kastrinos, Michael Glick, Adam J. Rosenbaum, Timothy Wang, and Alfred I. Neugut. 2011. The impact of suboptimal bowel preparation on adenoma miss rates and the factors associated with early repeat colonoscopy. *Gastrointest Endoscopy* 73, 6 (2011), 1207–14.

[17] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. 2016. LIRE: open source visual information retrieval. In *Proc. of MMSys*. Article no. 30.

[18] Jiquan Ngiam, Adam Coates, Ahbik Lahiri, Bobby Prochnow, Quoc V Le, and Andrew Y Ng. 2011. On optimization methods for deep learning. In *Proc. of ICML*. 265–272.

[19] R. Parmar, M. Martel, A. Rostom, and AN Barkun. 2016. Validated Scales for Colon Cleansing: A Systematic Review. *American journal of Gastroenterology* 111, 2 (2016), 197–204.

[20] 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, and Pål Halvorsen. 2017. Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proc. of MMSYS*.

[21] Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2008. Transfer learning for image classification with sparse prototype representations. In *Proc. of CVPR*. 1–8.

[22] 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, and Pål Halvorsen. 2016. Multimedia and Medicine: Teammates for Better Disease Detection and Survival. In *Proc. of ACM MM*. 968–977.

[23] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Thomas de Lange, Carsten Griwodz, Peter Thelin Schmidt, Sigrun L. Eskeland, and Dag Johansen. 2016. EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal Endoscopies. In *Proc. of CBMI*.

[24] Alaa Rostom and Emilie Jolicoeur. 2004. Validation of a new scale for the assessment of bowel preparation quality. *Gastrointestinal endoscopy* 59, 4 (2004), 482–486.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567* (2015).

[26] t. Rai, U. Navaneethan, T. Gohel, A. Podugu, P. N. Thota, R. P. Kiran, R. Lopez, and M. R. Sanaka. 2016. Effect of quality of bowel preparation on quality indicators of adenoma detection rates and colonoscopy completion rates. *Gastroenterol Report* 4, 2 (2016), 148–53.

[27] World Health Organization - International Agency for Research on Cancer. 2012. Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. (2012). https://goo.gl/IgZpVl Accessed: 2017-04-19.

Paper XIV

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

Deep Learning and Handcrafted Feature Based Approaches for Automatic Detection of Angiectasia*

Konstantin Pogorelov[1], Olga Ostroukhova[2], Andreas Petlund[3], Pål Halvorsen[1], Thomas de Lange[4], Håvard Nygaard Espeland[3], Tomas Kupka[3], Carsten Griwodz[1] and Michael Riegler[1]

***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 [1]. 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.**

***Index Terms*—Angiectasia, computer aided diagnosis, deep learning, machine learning, video capsular endoscopy**

I. Introduction

An obscure gastrointestinal (GI) bleeding is a common finding in the GI tract and caused by different diseases/conditions. The most challenging part is to detect the bleeding source in the small bowel either using video capsule endoscopes (VCEs) or via very invasive enteroscopy examinations. Superficial vascular lesions called angiectasia (see Figure 1 for an example) represent one of the most common source of bleeding in the small bowel and are therefore important to detect [2].

The most common procedure to detect angiectasia is to use VCEs. A VCE provides visualization of the GI tract by capturing images or recording a video by swallowing a pill-like disposable capsule equipped with one or more cameras. The camera pill contains a small processing device, a memory or wireless transmitter, and a battery. The VCE is swallowed by the patient, and it traverses and visualizes the GI tract for subsequent diagnosis and detection of GI diseases, such as angiectasia, by a doctor manually inspecting the video recordings. The latest generation of VCEs supports a maximum resolution of 520x520 pixels and is able to collect around 60,000 images per patient. Medical specialists assessing the images detect only around 69% of angiectasias (16% for the detection of bleeding to 69% for the detection of angiectasias) [1].

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

Fig. 1. 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 approach (c). Image taken from the GIANA dataset [3].

State-of-the-art software from the industry is able to reach an automatic detection sensitivity of 41% and a specificity of 67%. For a clinical scenario, this is clearly not reliable enough for automatic analysis. Both, sensitivity and specificity should be as close as possible to 100%, but at least larger than 85% for being used in a real clinical setting [1], [4]. Automatic detection of angiectasia is not very well researched, and there are only a few publications on the topic using saliency detection [5], [6]. However, no work has looked into machine learning using deep learning or handcrafted features. In this work, we therefore test different machine learning approaches to tackle automatic angiectasia detection in VCE videos. Using a publicly available and unbiased (equal number of negative and positive examples) dataset [3], we are testing algorithms (deep learning and handcrafted features-based) for frame-wise detection and pixel-wise localization. The best achieved results in this paper are 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.

The rest of the paper is organized as follows: first, we give an overview of the related work in the field. This is followed by a description of our methods, which we next experimentally evaluate. Finally, we conclude the paper and give directions for future work.

*This work is founded by the FRINATEK project "EONS" #231687

[1]Konstantin Pogorelov, Carsten Griwodz, Pål Halvorsen and Michael Riegler are with Simula Research Laboratory and University of Oslo, Norway konstantin@simula.no

[2]Olga Ostroukhova is with Research Institute of Multiprocessor Computation Systems n.a. A.V. Kalyaev, Taganrog, Russia

[3]Andreas Petlund, Håvard Nygaard Espeland and Tomas Kupka are with ForzaSys AS, Norway

[4]Thomas de Lange is with Institute of Clinical Medicine, University of Oslo, Norway

II. Related work

To the best of our knowledge, there exists almost no related work about automatic detection of angiectasia in VCE images or videos. The only known work is from Deeba et al. [5], [6] who present a saliency-based approach. The method is two-staged and unsupervised. In the first step, a patch distinctness (PD) map and an Index of Hemoglobin (IHb) map are created. In the second step, the PD and IHb maps are combined to create the final saliency map. On this final map, a local maximum search is performed to find regions-of-interests containing the lesions. In [5], a sensitivity of 100%, a specificity of 82.5% and an accuracy of 90.1% are reported on a dataset containing 50 normal images and 50 images containing angiectasia. In [6], the dataset is extended to 3,602 images with 968 containing angiectasia and 2,634 normal ones. The images are taken from 9 different videos, whereas 5 are from wired endoscopy and 4 from VCE. For the VCE videos, a sensitivity of 94.44% and a specificity of 83.92% are reported. The localization score is 95.04% and measures the fraction of correctly detected regions compared to the regions containing angiectasia.

Furthermore, since bleeding angiectasia looks quite similar to regular GI bleeding, a very common condition, work addressing automatic bleeding detection should be considered. For bleeding detection, a lot of related work exist, and the main challenge is that the bleedings do not occur in specific patterns, shapes, textures or colors, which makes them hard to detect. Furthermore, bleeding is usually caused by other intestinal diseases like angiectasia or cancer, etc [7]. The main methods to detect bleeding are based on handcrafted color and textural features. In [8], chrominance-moments-based texture and uniform local binary patterns in combination with a multi-layer perception neural network classifier are used to localize the source of bleeding in the VCE video. Methods working on pixel-level are shown to be more accurate to distinguish between bleeding and non-bleeding. Yuan [9] utilizes color features on pixel level of VCE frames and thresholds the color space to segment bleeding from normal mucosa. In [10], the authors perform super-pixel based segmentation to reduce the computational complexity and at the same time achieve high accuracy. In general, pixel level methods have higher accuracy than frame based, but are computationally more costly. This is an important factor taking into account that usually more than 60,000 frames have to be processed for one patient.

Most recent work is focusing on deep learning for bleeding detection by utilizing convolutional neural networks (CNN). In [11], the authors present a bleeding detection approach that uses CNNs. The presented CNN consists of eight layers and is basically a simple variation of the Imagenet architecture [12]. They report a recall/sensitivity of 99.2%, a precision of 99.9% and F1 score of 99.55%. For the training, 8,200 images were used, and 1,800 images were used for the testing. Both the training and testing dataset were biased towards the negative class. Furthermore, no cross validation for the evaluation was performed. Therefore, it cannot be ruled out that the almost perfect performance is based on overfitting. This was followed by another study using the same approach [13], but with a different dataset and on pixel accuracy level for segmentation. As the main metric, region intersection over union (IU) was used. For active bleeding, an IU of 0.7750 and for inactive bleeding an IU of 0.7524 were achieved. Another recent work that does not use deep learning but classifier fusion is the work by Deeba et al. [14] that combines two optimized Support Vector Machine (SVM) classifiers to detect bleeding. The features used by the classifiers are based on the RGB and HSV color spaces. For parameter tuning and evaluation, a cross validation approach is used, and they report an average accuracy of 95%, sensitivity of 94% and specificity of 95.3% for a dataset of 8,872 VCE frames.

The presented related work contains only two papers about angiectasia and some related work in the field of bleeding detection. As one can observe, even if deep learning is in the rise, handcrafted features still achieve good performance if used in a clever way. In the context of angiectasia, one can see that the VCE datasets used in the related work are biased and too small. Therefore, the goal of this work is to compare and evaluate deep learning and hand crafted features based approaches on a large and unbiased dataset.

III. Pixel-wise segmentation approach

The segmentation approach presented in this paper is able to pixel-accurate mark the angiectasia in the given frame. Based on our previous experience [15], [16], we decided to use generative adversarial network (GAN) to perform the segmentation. GANs [17] 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. We used a GAN model architecture initially developed for the retinal vessel segmentation in fundoscopic images called V-GAN as basis for our angiectasia segmentation approach. The V-GAN architecture [18] is designed for RGB images and provides a per-pixel image segmentation as output. To be able to use the V-GAN architecture in our angiectasia segmentation approach, we added an additional output layer to the generator network that implements an activation layer with a step function which is required to generate the binary segmentation output.

IV. Frame-wise detection approaches

Frame-wise detection approaches are designed to detect angiectasia on a frame level, i.e., if there is angiectasia in the frame or not. For frame-wise detection, we propose different methods where we conducted experiments using various configurations of our main methods. The main methods are global features (GF), deep features (DF) and a variation of our GAN approach. For the classification, we used the Random Tree (RT), Random Forrest (RF) and Logistic Model Tree (LMT) classifiers provided in the WEKA library [19].

Global features. For the GF method, we extracted handcrafted global features (describing the image on a global level, e.g., texture, color distribution, etc.) using the LIRE framework [20]. The features are Joint Composite Descriptor, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram and Pyramid Histogram of Oriented Gradients. We performed early fusion by combining all extracted features resulting in a feature vector with the size of 1186.

Deep features. For the DF approach, we used different well known working deep learning architectures to extract either the features directly (FEA) or to classify the images and using the whole range of concepts and their probabilities as input for the classifiers (CON). The architectures that we used are ResNet50 [21], VGG19 [22], and InceptionV3 [23].

Data augmentation. For fair performance comparison of the GF and DF approaches with the GAN approach, we implemented the same data augmentation (AUG) scheme (rotation and flipping of frames) as used in the training process of the GAN. Rotation was performed with 20° steps for the original and the flipped frames, resulting in 35 new frames complementary to the original ones.

GAN. The GAN detection approach utilizes a simple threshold activation function, which takes the number of positively marked pixels in the frame as an input. In the cross-validation experiments, we evaluated the activation thresholds from one pixel to a quarter of the frame. The best results were achieved with a threshold value of 2 pixels, which has been used for the detection experiments.

V. Experiments

The data used for all the experiments is from the GIANA 2017 challenge [3], 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 medical pathology diagnosis background. 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. For evaluation of the experiments, we used the precision (PREC), recall/sensitivity (SENS), specificity (SPEC), accuracy (ACC), F1 score (F1), Matthew correlation coefficient (MCC) and processing speed in number of frames per second (FPS) metrics. A detailed description and reasoning for the used metrics can be found in [24]. The localization metrics are calculated pixel-wise using the provided GT masks. For the best working approach (GAN), we also report detailed results for the ten-fold cross-validation including 95% confidence intervals (CI). For the detection part, we use a ZeroR classifier as baseline which assigns the label from the majority class (most common label in the dataset) to all the instances.

TABLE I
Ten-fold cross-validation results of the pixel-wise angiectasia areas the GAN segmentation approach.

Fold	PREC	SENS	SPEC	ACC	F1	MCC
1	0.805	0.877	0.999	0.999	0.839	0.839
2	0.893	0.908	0.999	0.999	0.901	0.900
3	0.870	0.871	0.999	0.999	0.871	0.870
4	0.808	0.884	0.999	0.998	0.844	0.844
5	0.876	0.894	0.999	0.999	0.885	0.885
6	0.838	0.849	0.999	0.998	0.843	0.842
7	0.900	0.887	0.999	0.999	0.893	0.893
8	0.863	0.900	0.999	0.999	0.881	0.880
9	0.866	0.914	0.999	0.999	0.889	0.889
10	0.873	0.817	0.999	0.999	0.844	0.844
95% CI	**0.859 ±0.020**	**0.880 ±0.018**	**0.999 ±0.001**	**0.999 ±0.001**	**0.869 ±0.015**	**0.869 ±0.015**

TABLE II
Ten-fold cross-validation results of the angiectasia frame-wise detection using the GAN approach.

Fold	PREC	SENS	SPEC	ACC	F1	MCC
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.000	0.967	1.000	0.983	0.983	0.967
3	1.000	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000	1.000	1.000
5	1.000	0.967	1.000	0.983	0.983	0.967
6	1.000	0.967	1.000	0.983	0.983	0.967
7	1.000	1.000	1.000	1.000	1.000	1.000
8	1.000	1.000	1.000	1.000	1.000	1.000
9	1.000	1.000	1.000	1.000	1.000	1.000
10	1.000	0.967	1.000	0.983	0.983	0.967
95% CI	**1.000 ±0**	**0.987 ±0.011**	**1.000** ±0	**0.993 ±0.005**	**0.993 ±0.005**	**0.987 ±0.011**

A. Results

Table I shows the results for the GAN localization algorithm (see Figure 1(b) and 1(c) for a comparison between the GT and the output of the GAN). 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 presented in Table II. The detection outperforms significantly the 85% requirements. Both result sets are a strong indicators that our GAN approach performs well for the tasks of angiectasia localization and detection. Finally, in Table III, 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. From 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

TABLE III
RESULTS FOR THE ANGIECTASIA FRAME-WISE DETECTION APPROACHES EVALUATED WITH THE ANNOTATED TEST SET.

Approach	PREC	SENS	SPEC	ACC	F1	MCC	FPS
GF+RT	0.570	0.568	0.568	0.568	0.566	0.138	130
GF+RF	0.628	0.623	0.623	0.623	0.620	0.252	105
GF+LMT	0.695	0.680	0.680	0.680	0.674	0.375	80
DF ResNet50 CON+RT	0.636	0.636	0.636	0.636	0.636	0.271	88
DF ResNet50 CON+RF	0.742	0.742	0.742	0.742	0.742	0.483	78
DF ResNet50 CON+LMT	0.734	0.732	0.732	0.732	0.731	0.465	53
DF ResNet50 FEA+RT	0.558	0.557	0.557	0.557	0.554	0.114	79
DF ResNet50 FEA+RF	0.721	0.720	0.720	0.720	0.720	0.441	70
DF ResNet50 FEA+LMT	0.748	0.738	0.738	0.738	0.736	0.486	46
DF VGG19 CON+RT	0.538	0.538	0.538	0.538	0.538	0.077	60
DF VGG19 CON+RF	0.594	0.593	0.593	0.593	0.592	0.187	49
DF VGG19 CON+LMT	0.545	0.545	0.545	0.545	0.544	0.090	32
DF VGG19 FEA+RT	0.515	0.515	0.515	0.515	0.515	0.030	54
DF VGG19 FEA+RF	0.548	0.548	0.548	0.548	0.548	0.097	47
DF VGG19 FEA+LMT	0.525	0.525	0.525	0.525	0.525	0.050	29
DF InceptionV3 CON+RT	0.537	0.537	0.537	0.537	0.537	0.073	66
DF InceptionV3 CON+RF	0.617	0.617	0.617	0.617	0.617	0.233	50
DF InceptionV3 CON+LMT	0.663	0.663	0.663	0.663	0.663	0.327	37
DF InceptionV3 FEA+RT	0.515	0.515	0.515	0.515	0.513	0.030	56
DF InceptionV3 FEA+RF	0.551	0.548	0.548	0.548	0.542	0.099	43
DF InceptionV3 FEA+LMT	0.533	0.533	0.533	0.533	0.533	0.067	30
AUG GF+RT	0.545	0.545	0.545	0.545	0.544	0.090	130
AUG GF+RF	0.650	0.643	0.643	0.643	0.639	0.293	105
AUG GF+LMT	0.627	0.625	0.625	0.625	0.624	0.252	80
AUG DF ResNet50 CON+RT	0.620	0.620	0.620	0.620	0.620	0.240	88
AUG DF ResNet50 CON+RF	0.787	0.787	0.787	0.787	0.787	0.574	78
AUG DF ResNet50 CON+LMT	0.765	0.763	0.763	0.763	0.763	0.529	53
AUG DF ResNet50 FEA+RT	0.553	0.553	0.553	0.553	0.553	0.107	79
AUG DF ResNet50 FEA+RF	0.727	0.723	0.723	0.723	0.722	0.450	70
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	-

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.

B. Conclusion

In this paper, we presented hand crafted and deep learning-based methods for automatic detection of angiectasia on a pixel- and frame-wise level. We compared several approaches (handcrafted and deep learning) and demonstrated, on a public available dataset, the capability of our proposed GAN approach to reach and exceed clinical requirements (sensitivity and specificity higher than 85%) for localization and detection performance. In summary, we achieved a sensitivity of 88% and a specificity of 99.9% for pixel-wise localization, and a sensitivity of 98% and a specificity of 100% for frame-wise detection. For future work, the improvement of the processing speed and verification with other pathologies for our best working approach is planed.

REFERENCES

[1] V. Baptista, N. Marya, A. Singh, A. Rupawala, B. Gondal, and D. Cave, "Continuing challenges in the diagnosis and management of obscure gastrointestinal bleeding," *World journal of gastrointestinal pathophysiology*, vol. 5, no. 4, p. 523, 2014.

[2] X. Bosch, E. Montori, M. Guerra-García, J. Costa-Rodríguez, M. H. Quintanilla, P. E. Tolosa-Chapasian, P. Moreno, N. Guasch, and A. López-Soto, "A comprehensive evaluation of the gastrointestinal tract in iron-deficiency anemia with predefined hemoglobin below 9mg/dl: A prospective cohort study," *Digestive and Liver Disease*, vol. 49, no. 4, pp. 417–426, 2017.

[3] J. Bernal and H. Aymeric, "Miccai endoscopic vision challenge angiodysplasia d & l," https://endovissub2017-giana.grand-challenge.org/home/, accessed: 2017-11-20.

[4] J. Regula, E. Wronska, and J. Pachlewski, "Vascular lesions of the gastrointestinal tract," *Best Practice & Research Clinical Gastroenterology*, vol. 22, no. 2, pp. 313–328, 2008.

[5] F. Deeba, S. K. Mohammed, F. M. Bui, and K. A. Wahid, "A saliency-based unsupervised method for angioectasia detection in capsule endoscopic images," *Proc. of the CMBES*, vol. 39, no. 1, 2016.

[6] ——, "A saliency-based unsupervised method for angiectasia detection in endoscopic video frames," *Journal of Medical and Biological Engineering*, pp. 1–11, 2017.

[7] M. Van Leerdam, E. Vreeburg, E. Rauws, A. Geraedts, J. Tijssen, J. Reitsma, and G. Tytgat, "Acute upper gi bleeding: did anything change?" *The American journal of gastroenterology*, vol. 98, no. 7, pp. 1494–1499, 2003.

[8] B. Li and M. Q.-H. Meng, "Computer-aided detection of bleeding regions for capsule endoscopy images," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1032–1039, 2009.

[9] Y. Yuan, B. Li, and M. Q.-H. Meng, "Bleeding frame and region detection in the wireless capsule endoscopy video," *IEEE Journal of biomedical and health informatics*, vol. 20, no. 2, pp. 624–630, 2016.

[10] Y. Fu, W. Zhang, M. Mandal, and M. Q.-H. Meng, "Computer-aided bleeding detection in wce video," *IEEE Journal of biomedical and health informatics*, vol. 18, no. 2, pp. 636–642, 2014.

[11] X. Jia and M. Q.-H. Meng, "A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images," in *Proc. of IEEE EMBC*, 2016, pp. 639–642.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[13] X. Jia and M. Q.-H. Meng, "A study on automated segmentation of blood regions in wireless capsule endoscopy images using fully convolutional networks," in *Proc. of IEEE ISBI*, 2017, pp. 179–182.

[14] F. Deeba, M. Islam, F. M. Bui, and K. A. Wahid, "Performance assessment of a bleeding detection algorithm for endoscopic video based on classifier fusion method and exhaustive feature selection," *Biomedical Signal Processing and Control*, vol. 40, pp. 415–424, 2018.

[15] 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*, pp. 1–33, 2017.

[16] 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.

[17] 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.

[18] 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.

[19] M. Hall, E. Frank, G. Holmes *et al.*, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[20] M. Lux, M. Riegler, P. Halvorsen, K. Pogorelov, and N. Anagnostopoulos, "Lire: open source visual information retrieval," in *Proc. of ACM MMSys*, 2016.

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

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

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. of IEEE CVPR*, 2016, pp. 2818–2826.

[24] 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, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. of ACM MMSYS*, 2017, pp. 164–169.

Paper XV

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

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

Konstantin Pogorelov[1,2], Olga Ostroukhova[7], Mattis Jeppsson[6], Håvard Espeland[6], Carsten Griwodz[1,2], Thomas de Lange[1,5], Dag Johansen[3], Michael Riegler[1,2,4] and Pål Halvorsen[1,2,4]

[1]*University of Oslo, Norway* [2]*Simula Research Laboratory, Norway* [3]*UiT - Artic University of Norway*
[4]*Simula Metropolitan Center for Digital Engineering, Norway*
[5]*Oslo University Hospital, Norway* [6]*ForzaSys AS, Norway*
[7]*Research Institute of Multiprocessor Computation Systems n.a. A.V. Kalyaev, Russia*

***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 [1]. 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 possible improvements in the work flow and the potential improved detection rates for medical experts.**

***Index Terms*—medical video analysis, machine learning, deep learning, image features, performance**

I. Introduction

Hospitals record and collect a huge amount of multimedia data which needs to be stored and analyzed, both on-the-fly and offline. One example is gastrointestinal (GI) tract examinations where large numbers of videos are collected, i.e., by an endoscope controlled by a medical expert. Making the future GI examinations more efficient and cost-effective is also a huge societal challenge as about 2.8 millions of new esophagus, stomach and colorectal cancers are detected yearly in the world with a mortality of about 65% [2]. All have a significant impact on the patients' health-related quality of life. Consequently, gastroenterology is one of the most significant medical branches. Colorectal cancer is the third most common cause of cancer mortality for both women and men, and it is a condition where early detection is important for survival. For example, a patient is going from a low 10-30% 5-year survival probability if detected in later stages of the disease to a high 90% survival probability in early stages [3].

Colonoscopy is considered to be the gold standard for the examination of the colon for early detection of cancer and precancerous pathology. However, it is not an ideal screening test. Polyps, which are abnormal growth of tissue projecting from a mucous membrane (see Figure 1(a)), are often predecessors of colorectal cancers, and are therefore important to detect early. However, on average, 20% of polyps are missed or incompletely removed, i.e., the risk of getting cancer largely depend on the endoscopists ability to detect polyps [1]. It is also a demanding procedure requiring a significant time investment from the medical professional, and the procedure is unpleasant and can cause great discomfort for the patient [4]. Furthermore, there are high costs related to the procedure. Norway has an average cost of about $450 per examination. In the US, colonoscopy is the most expensive cancer screening process with an average of $1,100 per examination, i.e., an annual cost of $10 billion dollars [5].

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

Fig. 1. Example of a polyp marked with a green circle (a), a corresponding polyp localization ground truth mask (b) and our output segmentation mask using GAN (c). Images taken from the CVC-968 [6].

In the area of image analysis and object detection, machine learning, and especially deep learning, has been very popular, also in the field of medicine, in the recent years. Deep learning algorithms are based on neural networks that use recently developed training techniques to train their models. They are basically an abstracted representation of data points. The representation is made on a high-level, and multiple layers for processing the networks are used to reach higher complexity. The different layers can learn different abstraction levels of the data using input of previous layers until they reach a final layer, which makes the final decision for the class. The new training techniques for deep learning were mainly made possible because of the emergence of GPU computing, which enables training of the networks in a reasonable amount of time. On the other side, the disadvantages include a very long training time, classification boundaries are hard to explain

(why one data point is put in this class), and they are very data driven [7], [8].

Automatic detection of polyps is in general well researched, and there are many publications on the topic. Related work indicates with a sensitivity and specificity close to one that the problem is solved. Nevertheless, there are still several open challenges, e.g., the evaluation of existing approaches is often performed on small and non-publicly available datasets. Medical datasets also have the challenge that they usually contain many true negative examples, but not so many true positives. Furthermore, a very important open question is how generalizable the proposed methods are. Generalization is a vital ability of a model trained on a dataset from one hospital to be applied in another hospital, e.g., using a different type of equipment (endoscope). Therefore, in this paper, we are addressing the challenges arising due to limited datasets and generalizability of models which both are common problems in medical multimedia scenarios [9].

The main contributions of this paper are proposing and testing different approaches to overcome the problems concerning generalization of models and limited datasets in terms of size and sample equality, and to propose a best approach for detection and localization of findings for medical image analysis. Our best working approach outperforms by far all our own and other tested approaches, and does at the same time not need a large amount of training data. Furthermore, it achieves good performance across datasets and does not need negative examples for training. With respect to dataset size and generalizability, we conclude that one proposed detection and localization model can be used across different datasets and different equipment and it is able to perform efficiently using very low amount of training samples. With our best working approach based on a generative adversarial network (GAN), we reach a detection specificity of 94% and an accuracy of 90.9% with only 356 training [6] and 6000 test [10] samples captured by different equipment.

The rest of the paper is organized as follows: first, we give an overview of the related work in the field. This is followed by a description of our methods, which we next experimentally evaluate. Finally, we conclude the paper and give directions for future work.

II. Related work

Recently, frame-wise detection and in-frame localization of colon polyps have been picked up as a research topic by many scientists in the medical imaging area, but lately also in the multimedia community. Approaches in context with automatic detection or localization of polyps in videos taken from colonoscopies can be divided into hand-crafted feature based, re-training or fine-tuning of existing and trained from scratch deep learning architectures.

In hand-crafted feature based approaches for detection, researchers extract features such as global or local image features (texture, edge or color based) from the frames and use them within different machine learning algorithms such as random forest (RF) or support vector machines (SVM) [11], [12]. The best working hand-crafted detection approaches are [13] and [9] with both precision and recall above 90%. The first approach [13] relies on edge and texture features whereas the latter [9] uses several different global image features. For localization, the best working approaches from Yuan et. al. [14], who use a bottom-up and top-down saliency approach, and from Wang et. al. [13], where they use edge and texture features. Usually, localization approaches can also be used for frame-wise detection.

Reusing already existing deep learning architectures and pre-trained models leads to very good results in for example the Imagenet classification tasks. Retraining architectures from scratch in the context of colonoscopies leads to reasonable good results, but the limited size of medical datasets is a problem for these approaches. For pre-trained models, even if their categories are quite different compared to the medical use case, it has been shown that they can be used in the context of polyp detection and localization tasks [15], [16], and that they often outperform hand-crafted approaches [17], [18].

In [19], a 3D convolutional neural network (CNN) architecture approach is presented for polyp detection. The method is also compared to hand-crafted and 2D CNN approaches, and it is shown that different approaches perform well for different sub-tasks. For example, the hand-crafted feature approach is working well for true negative detection. The best performance is reached with a fusion of all investigated approaches. Moreover, Pogorelov et. al. [20] and Riegler et. al. [21] compare different localization approaches (hand-crafted and deep learning). The conclusion is that pre-trained and fine-tuned deep learning models outperform other approaches, but that they are far away from being ready for clinical use (usually a sensitivity and specificity above 85% is considered as the borderline [22]).

In general, recent related work reports very promising results in terms of evaluation metrics, i.e., both recall (also called sensitivity) and specificity close to one. Nevertheless, most of the approaches are tested on small and non-publicly available datasets. Furthermore, the problem of medical datasets is that they usually contain many negative examples, but not so many positives is not well researched. Another open question is how generalizable the proposed methods are, meaning can a model trained on a dataset from one hospital be applied in another hospital. These are questions that we are addressing in this paper.

III. Methodology

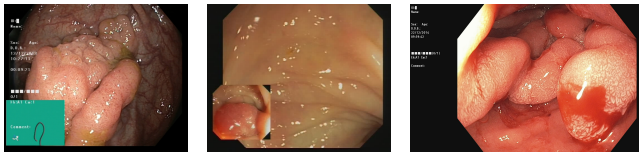
A. Pixel-wise segmentation/localization approach

The first presented segmentation approach is able to pixel accurate mark the polyp in the given frame. We use generative adversarial networks (GANs) to perform the segmentation. GANs [23] 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. We used a GAN model architecture initially developed for the retinal vessel segmentation in fundoscopic images, called V-GAN, as basis for our polyp segmentation approach.

The V-GAN architecture [24] is designed for RGB images and provides a per-pixel image segmentation as output. To be able to use the V-GAN architecture in our polyp segmentation approach, we added an additional output layer to the generator network that implements an activation layer with a step function which is required to 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 and to speed up the model training and data processing parts.

Data preparation. The frames used in this research is obtained from the standard endoscopic equipment and can contain some additional information fields related to the endoscopic procedure. Some types of the fields (see Figure 2), integrated into resulting frames showed to the doctor and captured by the recording system, can confuse detection and localization approaches, and it leads to frame miss-classification (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: 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 are excluded from further frame analysis.

Data augmentation. Due to a limited number of frames with the detailed ground truth masks, we implemented a data augmentation scheme used in the training process of the GAN. For the experiments presented here, we used only rotation and flipping of frames. Rotation was performed independently with 20° steps for the original. Together with the in-horizontal-direction-flipped frames, we added 35 new frames complementary to the original ones.



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

Fig. 2. 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 [6] and Kvasir [10].

B. Frame-wise detection approaches

Frame-wise detection approaches are designed to detect the target object on a per-frame level, i.e., in our GI scenario, detect if there is a polyp in the frame or not. For frame-wise detection, we propose different methods. We conducted experiments using various configurations of our main methods. The main methods are hand-crafted global features (GF-D), re-training and fine-tuning on existing deep learning architectures (RT-D) and a variation of the GAN approach (GAN-D) that was also used for the pixel-wise segmentation.

TABLE I
ARCHITECTURES AND CONFIGURATIONS USED FOR RT-D. WE HAVE USED THE *rmsprop* AND *SGD* OPTIMIZERS IN STEPS 1 AND 2, RESPECTIVELY, 50 EPOCHS AND A BATCH SIZE OF 32.

Method	Architecture	Step 2: frozen from layer	Image size
RT-D-Xcept	Xception [29]	26	299x299
RT-D-VGG19	VGG19 [30]	5	224x224
RT-D-ResNe	ResNet50 [31]	50	224x224

GF-D. For the GF method, we extracted handcrafted global features (describing the image on a global level, e.g., texture, color distribution, etc.) using the LIRE framework [25]. The features that we used are Joint Composite Descriptor, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram and Pyramid Histogram of Oriented Gradients. We performed early fusion by combining all extracted features resulting in a feature vector with the size of 1186. For the classification, we used a Logistic Model Tree (LMT) classifier from the Weka machine learning library [26].

RT-D. For the RT method, we implemented a re-training and fine-tuning approach and used it with three well known and working architectures. For all architectures, we used models trained on the Imagenet dataset for starting weights. The approach for RT-D works in two steps. First, we freeze all layers of the architecture and train only the base layers. After that, we unfreeze certain layers and fine-tune the network. Which blocks are un-freezed for the second step is decided via a Bayesian optimization algorithm [27] which runs for 20 iterations. To find good working optimizers, number of epochs and batch sizes for the different architectures, we also used Bayesian optimization for 20 iterations including all architectures. This lead to values that gave good overall results and could be used for all architectures to achieve better comparability. Details about the exact configurations and architectures used can be found in Table I. The dataset used for the optimization step is public available and details can be found in [28].

GAN. The GAN detection approach 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, which has been used for the detection experiments for the development and test set and confirms high performance of the GAN-based localization approach.

C. Block-wise segmentation/localization approach

The second localization approach is our attempt to utilize frame-wise 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 neighbor sub-frames. We have tested different window

TABLE II
OVERVIEW OF THE DATASETS USED IN THE EXPERIMENTS.

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

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 the 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.

IV. EXPERIMENTS

For the experiments, we use combinations of six different, publicly available datasets, namely CVC-356 [6], CVC-612 [32], CVC-968, CVC-12k [6], Kvasir [10] and parts of Nerthus [28] (see Table II for an overview). The CVC-356 and CVC-612 consist of 356 and 612 video frames, respectively. CVC-968 is a combination of CVC-356 and CVC-612. Each frame that contains a polyp comes with pixel-wise annotations in the CVC-356 and CVC-612 datasets. They are used for both training and testing in the localization performance evaluation experiments, and for the training only in the detection experiments. For the frame-wise detection approaches, except for the GAN-based approach, we also added the 1,350 class three frames with normal mucosa from the Nerthus dataset since normal mucosa examples for the negative class are required for our detection algorithms. The CVC-12k dataset contains 11,954 video frames. From these 11,954 frames, 10,025 contain a polyp and 1,929 show only normal mucosa. The polyps are not annotated pixel-wise, but with an oval shape covering the whole polyp (approximated annotation). For the Kvasir dataset, we included all classes except for the dyed classes (in a real world scenario something dyed is already detected by the doctor) leading to a dataset containing 1,000 frames with polyps, 5,000 without and only frame-wise annotations. The CVC-12k dataset is used as test set for block- and frame-wise detection and the Kvasir dataset for frame-wise detection.

A. Evaluation Metrics

For the evaluation of the experiments, we used the 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 can be found in [10]. The localization metrics are calculated pixel- and block-wise using the provided binary masks of the ground truth.

TABLE III
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.

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

B. Results

Table III depicts the performance evaluation results for the GAN-based pixel-wise segmentation approach. The best performance is achieved using the CVC-612 dataset for the 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 different datasets indicates that the approach works well, and the proposed localization algorithm can perform efficiently even with a low number of training samples available. This is important for our medical use-case scenario with a high diversity of objects and a limited amount of annotated data available. The initial localization experiments demonstrated more than 50% increase in performance of the localization using augmented training data, thus we have used augmented training data in all the pixel-wise localization experiments. A possible positive effect of test data augmentation with the following aggregation of the localization results will be subject of future research.

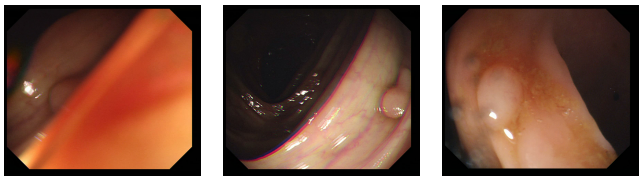
The results for the block-wise location approaches are presented in Table IV. The performance results obtained are especially interesting since all the approaches presented are trained with small amounts of training data without any negative examples (no normal mucosa frames at all). Furthermore, the CVC-12K dataset is heavily imbalanced which also 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. Furthermore, the algorithm used to combine the results on the different sub-frames into one can be improved by, for example, using another machine learning algorithm to learn the best combinations.

The frame-wise detection results can be found in Table V. 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

TABLE IV
PERFORMANCE OF THE BLOCK-WISE LOCALIZATION VIA DETECTION APPROACHES REPORTED PER METHOD AND USED TRAINING DATA.

Test set	Run	Training set	PREC	SENS	SPEC	ACC	F1	MCC
CVC-12k	LOC-Xcept-356	CVC-356	0.475	0.203	0.966	0.868	0.285	0.250
	LOC-Xcept-612	CVC-612	0.528	0.289	0.961	0.874	0.374	0.328
	LOC-Xcept-968	CVC-968	0.584	0.257	0.972	0.880	0.357	0.333
	LOC-VGG19-356	CVC-356	0.257	0.292	0.874	0.799	0.273	0.158
	LOC-VGG19-612	CVC-612	0.266	0.489	0.799	0.759	0.344	0.228
	LOC-VGG19-968	CVC-968	0.232	0.406	0.800	0.750	0.295	0.166
	LOC-ResNe-356	CVC-356	0.723	0.003	0.999	0.871	0.006	0.044
	LOC-ResNe-612	CVC-612	0.469	0.054	0.990	0.869	0.098	0.125
	LOC-ResNe-968	CVC-968	0.536	0.248	0.968	0.875	0.340	0.306

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 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 2 and 3).



(a) Blurry frame (b) Colors shift (c) Lens contamination

Fig. 3. 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 [6].

From the RT-D approaches, the Xcept has the best overall performance, and it performs best on the CVC-12k dataset. The ResNe method reaches best performance for the Kvasir dataset, but is still far away from the GAN approach (MCC 0.262 versus 0.689).

The GF-D 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 RT-D-VGG19. Overall, the RT-D 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 CNN called YOLOv2 [33]. 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 V
RESULTS FOR THE FRAME-WISE POLYP DETECTION APPROACHES. WE USED THE CVC-12K AND KVASIR DATASET AS INDEPENDENT TEST SETS.

Test set	Run	Training set	PREC	SENS	SPEC	ACC	F1	MCC
Kvasir	GAN-356	CVC-356	0.715	0.751	0.940	0.909	0.732	0.677
	GAN-612	CVC-612	0.595	0.803	0.891	0.876	0.684	0.619
	GAN-968	CVC-968	0.736	0.746	0.946	0.913	0.741	0.689
	GF-D-356	CVC-356	0.171	0.109	0.894	0.763	0.133	0.004
	GF-D-612	CVC-612	0.270	0.318	0.828	0.743	0.292	0.137
	GF-D-968	CVC-968	0.225	0.859	0.409	0.484	0.357	0.208
	RT-D-Xcept-356	CVC-356	0.358	0.259	0.907	0.799	0.300	0.190
	RT-D-Xcept-612	CVC-612	0.383	0.326	0.895	0.800	0.352	0.236
	RT-D-Xcept-968	CVC-968	0.459	0.256	0.939	0.825	0.328	0.251
	RT-D-VGG19-356	CVC-356	0.181	0.333	0.777	0.720	0.235	0.087
	RT-D-VGG19-612	CVC-612	0.213	0.583	0.682	0.669	0.313	0.186
	RT-D-VGG19-968	CVC-968	0.231	0.320	0.842	0.774	0.268	0.142
	RT-D-ResNe-356	CVC-356	0.236	0.178	0.885	0.767	0.203	0.070
	RT-D-ResNe-612	CVC-612	0.321	0.507	0.785	0.739	0.393	0.247
	RT-D-ResNe-968	CVC-968	0.248	0.877	0.469	0.537	0.387	0.262
	YOLO-968	CVC-968	0.530	0.559	0.901	0.844	0.544	0.450
CVC-12k	GAN-356	CVC-356	0.967	0.624	0.888	0.667	0.758	0.378
	GAN-612	CVC-612	0.934	0.609	0.778	0.636	0.737	0.286
	GAN-968	CVC-968	0.906	0.912	0.510	0.847	0.909	0.428
	GF-D-356	CVC-356	0.829	0.909	0.030	0.767	0.867	-0.081
	GF-D-612	CVC-612	0.809	0.383	0.530	0.407	0.520	-0.064
	GF-D-968	CVC-968	0.835	0.854	0.125	0.737	0.845	-0.020
	RT-D-Xcept-356	CVC-356	0.913	0.624	0.693	0.636	0.742	0.236
	RT-D-Xcept-612	CVC-612	0.876	0.740	0.457	0.694	0.802	0.160
	RT-D-Xcept-968	CVC-968	0.899	0.690	0.600	0.676	0.781	0.224
	RT-D-VGG19-356	CVC-356	0.257	0.292	0.874	0.799	0.273	0.158
	RT-D-VGG19-612	CVC-612	0.266	0.489	0.799	0.759	0.344	0.228
	RT-D-VGG19-968	CVC-968	0.232	0.406	0.800	0.750	0.295	0.166
	RT-D-ResNe-356	CVC-356	0.723	0.003	0.999	0.871	0.006	0.044
	RT-D-ResNe-612	CVC-612	0.232	0.406	0.800	0.750	0.295	0.166
	RT-D-ResNe-968	CVC-968	0.870	0.303	0.766	0.378	0.450	0.057
	YOLO-968	CVC-968	0.932	0.641	0.757	0.660	0.759	0.296

V. CONCLUSIONS

In this paper, we have presented hand crafted and deep learning-based methods for automatic, pixel-, block- and frame-wise detection of polyps in videos from colonoscopies.

We evaluated the performance of our methods on different datasets. To achieve real-world comparability, we chose difficult datasets captured using different hardware equipment that were imbalanced in terms of positive, and negative examples and we also performed performance validation using different datasets for training and testing. Additionally, we tried to use as little amount of training data as possible. We showed that our newly proposed GAN based method outperforms handcrafted features and approaches based on well-known and working deep learning architectures. With our best working GAN-based approach, we reached detection specificity of 94% and accuracy of 90.9% with only 356 training and 6,000 test samples for the data captured by different equipment in different hospitals. The localization specificity and accuracy for the same training set are 98.4% and 94.6% respectively. Thus we can conclude that our approach works with a little amount of training data and, moreover, does not require negative examples for training, which is important to be able to use lesion imagery, already collected in hospitals. For future work, we plan to improve all methods presented in this paper with the main focus on the GAN-based approach, extend the experiments to other datasets and compare it to a broader range of approaches including a time-series-based analysis using for example long short-term memory.

REFERENCES

[1] 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*, vol. 362, no. 19, pp. 1795–1803, 2010.

[2] World Health Organization - International Agency for Research on Cancer, "Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012," 2012.

[3] H. Brenner, M. Kloor, and C. P. Pox, "Colorectal cancer," *Lancet*, vol. 383, no. 9927, pp. 1490–502, 2014.

[4] 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*, vol. 101, no. 2, pp. 343–350, 2006.

[5] The New York Times, "The $2.7 Trillion Medical Bill," http://goo.gl/CuFyFJ, [last visited, Nov. 29, 2015].

[6] J. Bernal and H. Aymeric, "Miccai endoscopic vision challenge polyp detection and segmentation," https://endovissub2017-giana.grand-challenge.org/home/, accessed: 2017-12-11.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.

[9] M. Riegler, C. Griwodz, C. Spampinato, T. de Lange, S. L. Eskeland, K. Pogorelov, W. Tavanapong, P. T. Schmidt, C. Gurrin, D. Johansen, H. Johansen, and P. Halvorsen, "Multimedia and medicine: Teammates for better disease detection and survival," in *Proc. of ACM MM*, 2016, pp. 968–977.

[10] 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, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. of MMSYS*, june 2017, pp. 164–169.

[11] A. Mamonov, I. Figueiredo, P. Figueiredo, and Y.-H. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1488–1502, July 2014.

[12] 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," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1379–1389, 2014.

[13] 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*, vol. 120, no. 3, pp. 164–179, 2015.

[14] Y. Yuan, D. Li, and M. Q.-H. Meng, "Automatic polyp detection via a novel unified bottom-up and top-down saliency approach," *IEEE Journal of Biomedical and Health Informatics*, 2017.

[15] J. Bernal, N. Tajkbaksh, F. J. Sánchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge," *IEEE Transactions on Medical Imaging*, vol. 36, no. 6, pp. 1231–1249, 2017.

[16] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[17] Y. Shin and I. Balasingham, "Comparison of hand-craft feature based svm and cnn based deep learning framework for automatic polyp classification," in *Proc. of EMBC*, 2017, pp. 3277–3280.

[18] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Proc. of IEEE ISBI*, 2015, pp. 79–83.

[19] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 65–75, 2017.

[20] 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*, vol. 76, no. 21, pp. 22 493–22 525, 2017.

[21] 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)*, vol. 13, no. 3, p. 26, 2017.

[22] V. Baptista, N. Marya, A. Singh, A. Rupawala, B. Gondal, and D. Cave, "Continuing challenges in the diagnosis and management of obscure gastrointestinal bleeding," *World journal of gastrointestinal pathophysiology*, vol. 5, no. 4, p. 523, 2014.

[23] 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.

[24] 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.

[25] M. Lux, M. Riegler, P. Halvorsen, K. Pogorelov, and N. Anagnostopoulos, "Lire: open source visual information retrieval," in *Proc. of ACM MMSys*, 2016.

[26] 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*, vol. 11, no. 1, pp. 10–18, 2009.

[27] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[28] K. Pogorelov, K. R. Randel, T. de Lange, S. L. Eskeland, C. Griwodz, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Nerthus: A bowel preparation quality video dataset," in *Proc. of MMSYS*, june 2017, pp. 170–174.

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint arXiv:1610.02357*, 2016.

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

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

[32] 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*, vol. 43, pp. 99–111, 2015.

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