# EIR - A Medical Multimedia System for Efficient Computer Aided Diagnosis

Michael Alexander Riegler

13.10.2016

## Abstract

Health care systems all over the world have a long history of adopting technology for being able to improve care, quality of life and patient survival. Visual information is frequently used to support medical experts in their daily tasks, such as disease detection and analysis. In this context, computer vision and medical imaging are important tools to provide essential support. Furthermore, multimedia data produced by health care systems is growing, and it is a general misconception that disease detection and assessment are provided exclusively by computer vision and medical imaging. Core competences 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 can play an important role for the successful improvement of health care systems. The societal impact that multimedia research can have by addressing challenges and open problems in the field of medicine should therefore not be underestimated.

As a first attempt to address this challenge, our work explores different fields in multimedia research, starting from annotation of multimedia content over automatic analysis of the content and efficient processing of the workloads to visualization. Therefore, we have researched and developed a medical multimedia system addressing a use case with an important societal impact. This use case is disease detection in the gastrointestinal (GI) tract of the human body to be able to support medical experts in their work. The early 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 and is a non trivial task.

Investigating the field of GI diseases from a multimedia research point of view required several steps of research and development. First, we looked into a search-based classification algorithm, which serves as the basis of the system. The developed algorithm is based on information retrieval methods and has the advantage of being both fast and accurate at the same time. We also tested the algorithm on different use cases to demonstrate flexibility and accuracy before we applied it to our medical scenario. Next, we created a complete medical multimedia system called EIR, after the Norwegian goddess of medicine and healing. We researched and developed different subsystems for the EIR system. These subsystems are (i) the annotation subsystem, which makes it possible to collect data and transfer knowledge from the medical experts into our system; (ii) the detection and automatic analysis subsystem, which is responsible to detect and analyze the medical multimedia content automatically; and (iii) the visualization subsystem that provides the information to the medical personnel.

Furthermore, the focus of the EIR system lies on accurate and time efficient processing of multimedia data within the system. We investigated therefore parallel processing, GPU-based acceleration and different classification approaches that are evaluated and compared with state-of-the-art methods such as deep learning.

We demonstrated that the EIR system can outperform state-of-the-art approaches in both processing speed and detection accuracy reaching detection accuracy above $90\%$ and processing speed above $300$ frames per second. With our good results we could attract several hospitals for collaboration, and the EIR system is momentarily prepared for being tested and used under clinical conditions within our collaborating hospitals in Norway, Italy, Japan, Sweden and USA.

# Acknowledgements

First of all, I would like to thank my two official and two unofficial supervisors; Pål Halvorsen, Carsten Griwodz, Martha Larson and Mathias Lux. Without their support and motivation this work would not have been possible. Martha once told me that it is very important who is your "Doktorvater". I did not really understand the true meaning at the time but I do now and I would like to extend her advice: "The people that believe and support you from the begin of your scientific journey (masters) to the PhD are extremely important throughout the journey to reach this milestone". I can say without any doubt that I am what I am today because of the people I met on this journey.

In particular, I would like to thank; Pål for the possibility to do my PhD at Simula even if I was not the perfect fit for the position, his endless endurance getting used to my strange research topics and basically giving me total freedom in the work although still guiding me through it. Martha for her never ending and encouraging enthusiasm, challenging problems and always lending an ear helping out with problems appearing along the way. Carsten for his challenging and motivating discussions and feedback - often disastrous to the work at the beginning, though ending up a lot better. Mathias for his crazy ideas, motivation and teaching me about design and how to spice up research.

I would also like to give many thanks to all my current and former colleagues in the Media Performance Group, taking me in as one of their own even if I was a bit of an alien from a research interest point of view. I will not write individual text to any of you because you are all equal important, thank you Vamsi, Jonas, Konstantin, Iffat, Andreas, Håkon, David, Minoo, Olga, Ragnhild, Kjetil, Preben, Lilian and Robin. Next I would like to thank the Multimedia Computing Group at the Technical University of Delft, especially; Raynor Vliegendhart, Yue Shi, Alessio Bazzica, Xinchao Li, Christoph Kofler, Wen Li, Babak Loni, Cynthia Liem, Alan Hanjalic, Carsten Eickhoff and last but not least Victoria from who I learned several important lessons.

A huge thank you also goes to Tien for being both a very good friend and fellow researcher, always having an open ear for me and with whom I hopefully will collaborate with through many more years! I would also like to thank all my co-authors and collaborators. For me, one of the most interesting and motivating parts of being a researcher is the possibility to work with different people, creating cool things. I was lucky to met a lot of cool people and together we performed a lot of interesting research. Special thanks here to Thomas de Lange and Sigrun Losada Eskeland for showing interest in our work even if it was totally out of their medical domain. A big thank you also to Herwig and Gerda who are my link to Austria, for always supporting me and sending survival packs for hard times!

Finally, I would like to thank my family Sunniva, her family and my beloved son Ask.

Especially Sunniva for her love, motivation and support and Ask for being the coolest baby, for not being born before my submission deadline and being partially the motivation for finishing fast.

In the case I forgot to mention someone (I hope I did not) a big thank you to everyone I forgot to mention. As an excuse, I was very tired when I wrote this. I really wonder where the saying "I slept like a baby" comes from, they truly do not sleep that heavy or much... Anyway, thanks a lot and:

MAY THE FORCE BE WITH YOU

THE PROFESSOR & THE GRIFF

# Contents

# List of Figures

# List of Tables

# Part I

# Overview

# Chapter 1

# Introduction

A huge part of today's life is related to multimedia content and also the health care system produces more and more multimedia content. The estimated size of data in the health care system for the whole world is around $162$ exabyte with an estimated increase of $2.5$ exabyte per year [11]. One can see that the amount of data that is created by the medical field will in the future reach gigantic scales [142], which comes with several challenges, like how to analyze, store or transmit it.

Before processing health related multimedia data to, for example, support medical doctors, a very important but also challenging aspect is the understanding, analysis and deployment of the content. When it comes to consumption by the medical users, like in video streaming based patient examination, communication or other medical tasks, the time dimension regarding speed and near real-time representation is important. Another problem that comes naturally with a large amount of data is how to find data efficiently and how to make such an immense amount of data retrievable. Because of the large amount of multimedia data in the health care system, parallel processing and elastic heterogeneous resources are important to achieve timing support for multimedia workloads by being able to process a large amount of data in parallel at the same time. In this work, we investigate how multimedia workloads in the medical field can be efficiently and automatically analyzed to support medical experts in their tasks.

Since the medical field by itself is huge, we decided to specifically address one area in this field. We decided for the human gastrointestinal (GI) system (figure 1.1) 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 [192].

An accurate automatic medical analysis system will have high impacts in the medical sector influencing patient survival rates, clinical work flows and costs. In the GI field, medical imaging has created visual representations of the interior of a body. However, to automatically detect and locate diseases, image representations are not sufficient. There is a need for image and video processing, analysis, information search and retrieval, combination with other sensor data and medical experts in the loop, and it all needs integration and efficient processing [142].

This work contributes to this by investigating efficient analysis and processing of multimedia workloads in the field of GI endoscopy with the goal to research new methods and in the best case to create a complete prototype of a medical multimedia system.

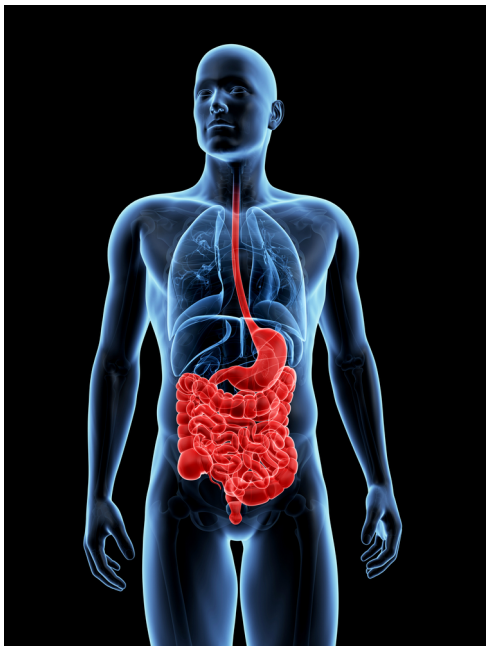


Figure 1.1: A complete overview of the human GI tract (shutterstock.com).

## 1.1 Background and Motivation

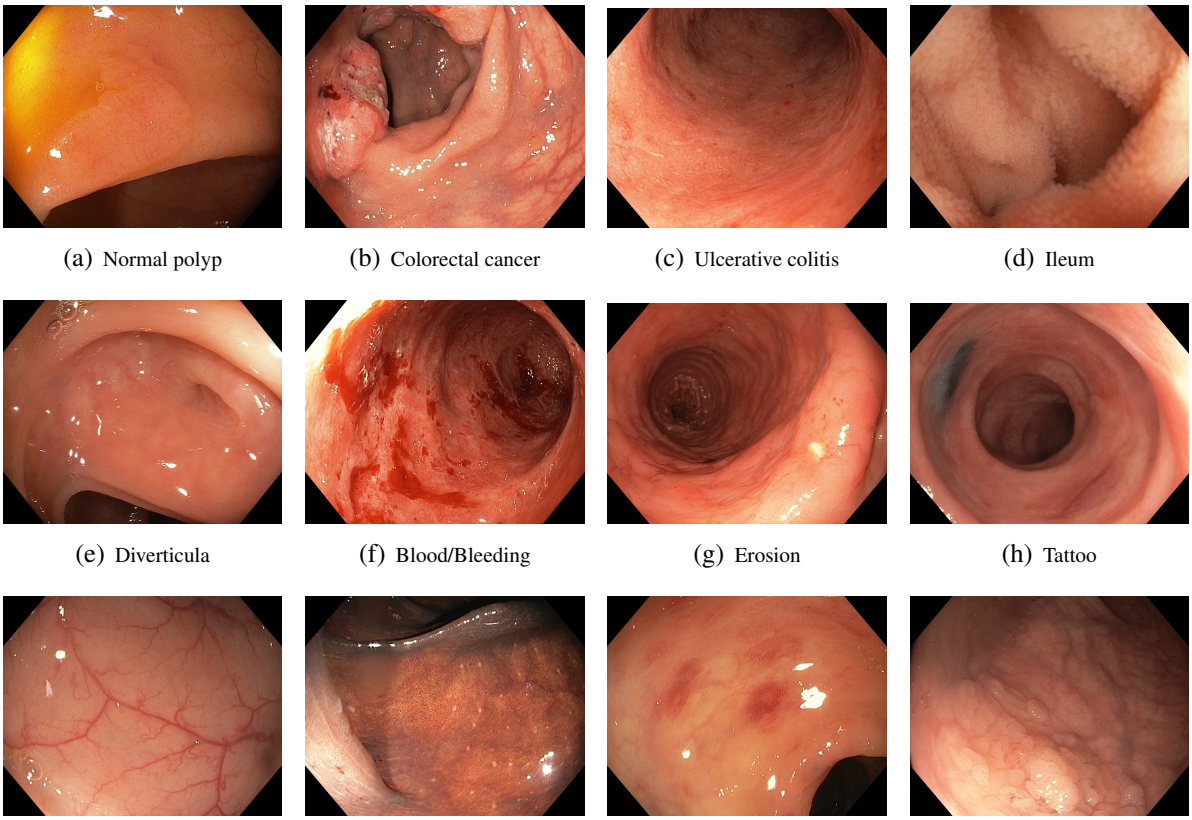
The human digestive system can potentially be affected by many types of diseases. For example, three of the six most common cancer types are located in the GI tract, with about 2.8 million new luminal[1] GI cancers (esophagus, stomach, colorectal) yearly and a mortality of about 65% [192]. Early detection is critical for the outcome, level of treatment and survival of the patients. Common diseases in the GI tract, beside colorectal cancer (CRC), include gastroesophageal reflux disease, peptic ulcer disease, inflammatory bowel disease, celiac disease (genetic autoimmune disorder that leads to problems with the digestion of gluten food) and chronic infections. Some visual examples of the most common diseases can be found in figure 1.2. The examples in the images range from a normal polyp to advanced polyps.

As Verdens Gang (a Norwegian Newspaper) describes (figure 1.5), Norway has one of the highest incidences of CRC worldwide and the numbers are increasing. All possible GI diseases have a significant impact on the patients' health-related quality of life [18]. Consequently, gastroenterology is one of the biggest medical branches.

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 example of an endoscope used for such an examination can be found in figure 1.3. An alternative to the manual colonoscopy is to perform the examination using a camera pill, which is a wireless capsular video endoscope (VCE) that can be swallowed by the patient and is able to record a video from the whole GI system. An example for a VCE device is shown in figure 1.4.

However, scheduled testing (screening) of a population for a complete country is challenging due to high costs, limited willingness by the patients to undertake the unpleasant procedure, high time consumption for the medical experts and shortage of qualified medical personnel.

[1] A structure inside the space of a tube-like structure. For the human body, this can for example be the nasal tract or intestines.



(a) Normal polyp (b) Colorectal cancer (c) Ulcerative colitis (d) Ileum

(e) Diverticula (f) Blood/Bleeding (g) Erosion (h) Tattoo

(i) Normal Vessel (j) Melanosis (k) Erythema (l) Advanced Polyps

Figure 1.2: An inconclusive list of diseases that can be diagnosed using colonoscopy [139].

Early detection is critical for the outcome, level of treatment and survival. Screening for CRC on a regular basis is therefore recommended. Statistics show, for instance, that the life-time risk of getting CRC, the second most common cancer for both genders, is 6% [38].

However, colonoscopy (the endoscopic examination of the colon) is unpleasant for the patients, each requires about two man-hours of medical personnel and often lesions are missed because of tiredness of the medical doctor or a specific part in the colon was not reachable due to narrow passages in the colon.

In the USA, colonoscopy is the most expensive 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) [172, 173]. In the UK, the costs are around $\$2,700$ per person [153]. In Norway, screening costs around $\$1,600$ per patient [114]. Obviously, scaling this to a population-sized cohort[2] is very resource demanding and incurs large costs.

Hospitals record, store and process large amounts of data. However, the collected data is not used efficiently 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. This is emphasized by the following comment from a medical doctor from one of our partner hospitals:

*"I have a lot of data lying around. Like for example images, videos, sensor logs, patient records and so on. Unfortunately, I am not able to use all the different types of data like I would*

[2]A cohort is a subset of people with a shared characteristic. For CRC screening in Norway this would be everyone above 50 years.


Figure 1.3: A colonoscopy endoscope ready for the examination of a patient.

*like to do. They are just stored on a computer somewhere. I even don't know where, and I don't think the IT support really know 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."*
*– A doctor in the Vestre Viken Hospital in Norway, September 2015 [139].*

After collecting this and many similar statements about not enough time for manual analysis, we teamed up with specialists in the area of GI diseases to investigate how multimedia research can improve medical systems. We soon detected that the multimedia data in the medical field has huge but not used potential. In this work, we discuss why multimedia researchers are needed in the medical field, why medical image processing alone is not the key to solve their challenges and we also present such a medical multimedia system built for the GI endoscopy use case.

An international cooperation of computer science researchers, medical experts and manufacturers of medical equipment has been established and will also continue the work after this PhD. The main goal is an automated detection and interpretation of lesions and diseases in the GI tract and subsequently remedy the shortage of qualified medical personnel by computerizing and automating some of the most complex and labor-intensive task with the help of multimedia methods and technology.

## 1.2 Problem Statement

To aid and scale GI tract examinations, we have started inter-disciplinary research of a multimedia system, called EIR after the Norwegian goddess of healing, which supports endoscopists in the detection and interpretation of diseases in the entire GI tract. The overall goal is to develop both, (i) a live system assisting the visual detection of diseases during colonoscopies that is verified with different use cases, and (ii) a future fully automated screening for the GI

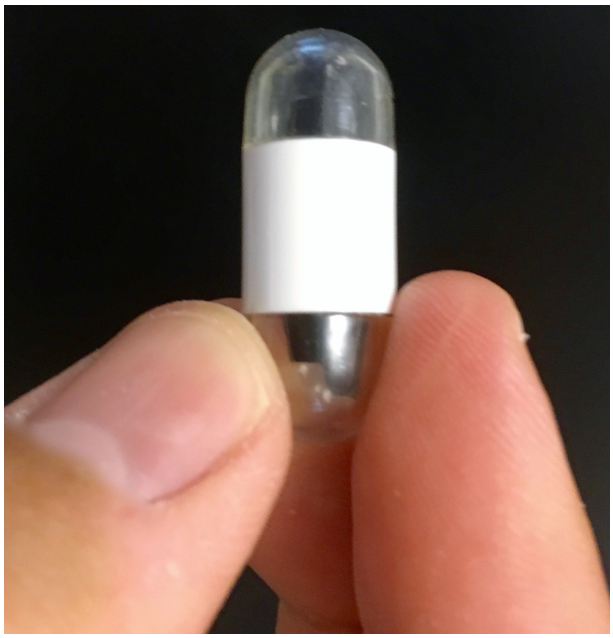


Figure 1.4: A wireless capsular video endoscope ready to be swallowed by a patient.

tract using VCE, i.e., a small capsule type device with an image sensor. These aims come with strict requirements on 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 process at least 25-30 frames or images per second (FPS)) [139].

**The research question for this thesis is:** ***How can multimedia data in our GI tract use case efficiently be exploited to support the medical experts in disease detection and live patient examinations?*** The goal of this thesis is to be a first cornerstone for building a complete multimedia system that can help to answer parts of our research question to achieve the overall goal and have societal impact by helping people to survive lethal diseases. From our question, we define the problems targeted by this thesis as following:

**Main Objective:** 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 of pathological endoscopic findings and anatomical landmarks in the GI tract. The system should include the whole pipeline from content creation and annotation, over 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:** Research and develop a subsystem that can be used by the medical doctors to annotate videos or images efficiently. Such an annotation tool has to be easy to use and understand by the medical experts. Furthermore, it should be designed in a way that it can help them to minimize the amount of time that they have to invest for the annotation task.

**Sub-objective 2:** Research and develop a subsystem for computer-based detection and decision supported for live endoscopies and VCEs. During the live colonoscopy performed

# Tarmkreft - den tause folkesykdommen

**Av: Marie Moen Kingsrød**

Over 4000 nordmenn årlig, og stadig flere, får tarmkreft. Vi ligger på verdenstoppen. Mange overlever ikke. Noen fordi de går for sent til legen,

Figure 1.5: A Norwegian newspaper article about the danger of colon cancer in Norway. It describes that colon cancer is an often overlooked problem and that every year more than $4,000$ Norwegians get infected by it [102].

Figure 1.6: A colonoscopy room with a colonoscope ready for patient examination [139].

with the equipment shown in figure 1.6, the video should also be analyzed by our system (as shown in figure 1.8) for computer-assisted detection and localization, giving the clinicians a signal if an endoscopic finding is detected. Furthermore, the system should fully automatically analyze videos recorded by VCEs (figure 1.7). Moreover, enable future large scale first level automatic screening, an easier (home-based) access and increased participation due to decreased discomfort for the patients.

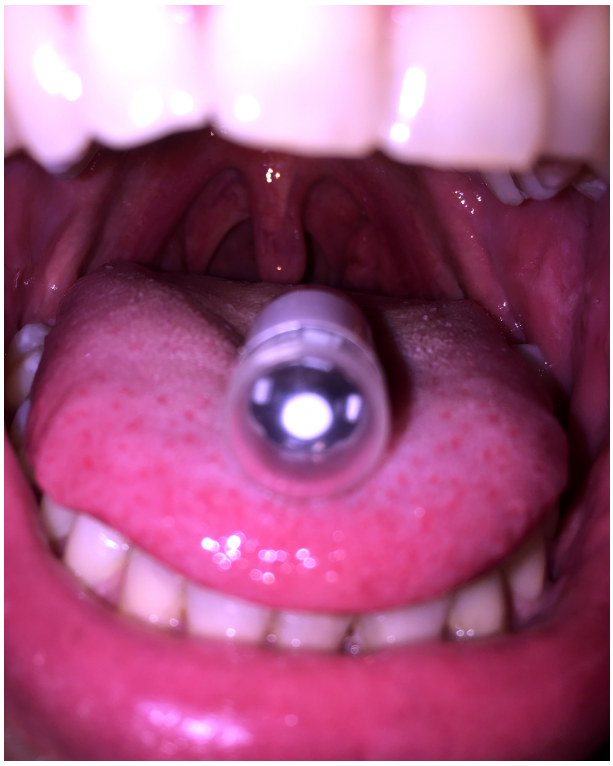


Figure 1.7: A patient swallowing a wireless video capsular endoscope (VCE) [139].

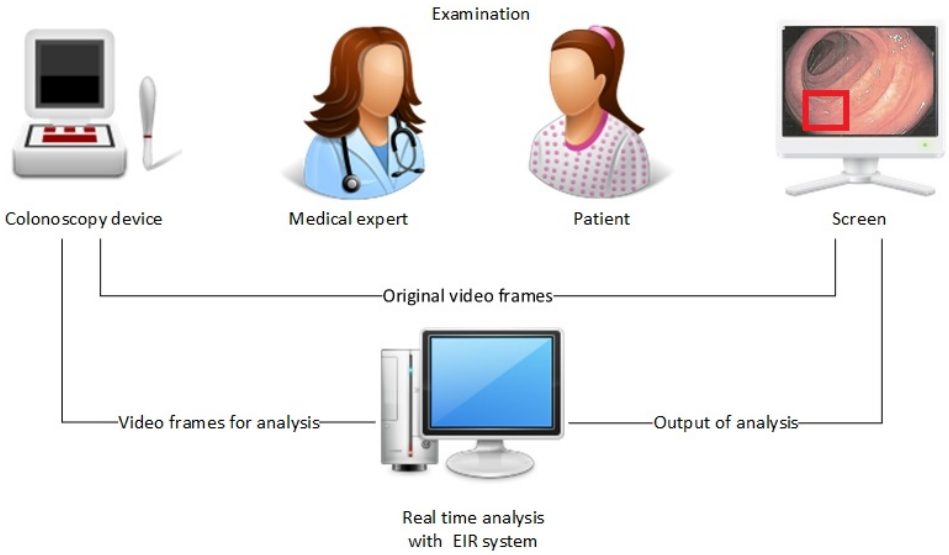



Figure 1.8: An overview of the planed live system. The video frames are captured directly from the colonoscopy device, analyzed and presented on screen to the doctor in real-time.

**Sub-objective 3:** Visualization of the automatic findings from examinations in an own subsystem, i.e., decrease time spent by medical personnel in reporting. Furthermore, the subsystem should provide the automatic results and the collected data for institutions in an easy to use way that can be used to support training of gastroenterologists.

## 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 disease detection in the GI tract, using mainly video data but prepare the system for the usage of different data types.

We decided to limit the focus of this initial phase of our larger project to polyp detection in the GI tract using mainly videos because of two main challenges. The first one is the large number of possible diseases and their different visual appearances in the GI tract. The second one is the lack of publicly available data for different diseases, which makes it hard to evaluate and compare methods. At least for the polyp use case their are some datasets available that can be used for the evaluation of the methods researched and developed for this work.

Other limiting factors are the contrasting types of equipment used in different hospitals and how the data is collected by the medical doctors. For example, doctors in our partner hospitals in ASU Mayo, Vestre Viken Hospital Trust, Rikshospitalet and the Karolinska University Hospital use different endoscopes for collecting their videos. This leads to videos and images with different attributes like different resolutions or light conditions. Furthermore, they have different standards for their examinations, for example, when to take a picture or a video or how long the recorded videos should be. In Karolinska, for example, 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. To make the system useful for the different hospitals and configurations, it should be modifiable based on each of them, which in our case means that it should be trainable with the individual data from each hospital if necessary.

It is important to point out that even with our focus to the polyp use case, the system is built in a way that it can easily be extended with different diseases if training data is available. As additional scopes, we define that the system should be verified with different use cases to show that it is generalizable, and that it supports efficient processing in terms of time and amount of data.

## 1.4 Research Methods

The ACM Education Board created in 1989 by a task force on the core of computer science, determines and characterizes the structure of how research in computing, should be approached. This report [30] defines computer science in its essence as an intersection between several central processes. The central processes are applied mathematics, science and engineering. These central processes are basically reflected in the paradigms of (i) theory, (ii) abstraction and (iii) design.

For this thesis, we worked a lot on practical system aspects relating most to the design process. This has been done by creating prototypes for a disease detection system that can be used in hospitals. By working close with medical personal like the Vestre Viken Hospital Trust and the Karolinska University Hospital, we gained insights about domain specific requirements and knowledge. Nevertheless, the thesis touches on the elements of all three processes. The following gives and overview how the thesis fits within each process.

- **Theory:** The theory process is responsible for defining and characterizing the objects under study by formulating and hypothesize possible relationships. Furthermore, it is characterized by determining relationships among objects, verifying their correctness and interpreting the results.

  For the theoretical part, we touched upon elements of image processing in 2D geometries, human interpretation of multimedia content and linear algebra, etc. In the design of the algorithmic basic for the system, we developed a search-based classification algorithm and as one of the use cases the theory of intentional framing of images.

- **Abstraction:** The abstraction process is used for modeling and emerges from experimental scientific methods. While a researcher is investigating a problem, a hypothesis is formed, a model created, experiments designed and finally data collected and analyzed.

  We use experiments and different datasets to verify our hypothesizes. We also apply it to different use cases to show that is has potential to be generalized and applied to other problems than the medical use case. We explore image retrieval techniques for classification. We perform several image and multimedia data processing operations in the different use cases. Furthermore, we study the performance of our system in terms of accuracy and speed. For some of the use cases, we also study the users response to our solution. To achieve this, we often used crowdsourcing and designed several user studies.

- **Design:** The last process is design, which is closely related to engineering. This involves researchers to state requirements and solutions, followed by designing and implementing a system. This process is concluded with and evaluation of the system.

  All the theories and abstractions presented in the thesis are actually implemented in a real world system and tested with real world data such as user feedback. The system uses multiple architectures and devices. Parts of the system are tested and used by medical personnel to verify if they are useful or not.

## 1.5 Contributions

The summarized main contributions of this thesis are:

- (i) Technical development of a medical multimedia system called EIR including annotation, detection and visualization tools that demonstrates the potential of multimedia research in the health care system.

- (ii) Develop an efficient, generalizable content-based method to process multimedia data.

- (iii) Research how distributed processing can help to achieve real-time performance for medical multimedia workload processing.

- (iv) Show why the multimedia community should apply their research in medicine, and illustrate how multimedia technology and methods can be used in the medical field to improve work flows, patient care and most important possible save lives.

- (v) Implementation and presentation of prototypes and demos of the system and making parts of it open source.

- (vi) Writing and publishing several research papers about our findings and experiences and share it with the community.

All main contributions of the thesis are supported by publications in top tier conferences or journals. The diagram in figure 1.9 gives an overview about which of the attached papers contribute to which objectives. In more detail, the main contributions in coherence to the objectives defined in section 1.2 of the thesis are:

- **Contributions to the main objective:** We developed the EIR system for automatic detection of lesions in the GI tract. The system consists of an annotation, a detection and localization and a visualization subsystem. This 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 [142, 143, 127, 145, 126].

  Using the ASU Mayo dataset [168], we showed that EIR reaches high performance in terms of accuracy and processing. For the classification part, we can report a sensitivity of almost 98% and a precision of almost $94\%$. This means that EIR is able to find polyps in almost all cases with a high precision. This can help the medical experts to save time and lives [142, 143, 127, 145, 126]. We could also show that the EIR system is able to perform multi-class classification and that the search-based Multiclass-EIR approach is able to outperform Deep-EIR, which is based on state-of-the-art deep learning techniques. Nevertheless, it is important to point out that the used dataset is limited in size and that evaluations on larger amount of data are recommended as soon as the data is available.

  Moreover, we compared EIR with other existing 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 [142, 126, 145].

  For each part of the EIR system, we developed working prototypes and demo applications. These prototypes and demo applications have been presented at conferences [4, 142, 126, 145].

  For the real-time processing challenge, we showed that EIR 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 [142]. We researched and implemented different ways of distributed and parallel processing by using different architectures to improve the performance of the EIR system. One of the methods that we researched is the distribution of the detection and localization part on graphics processing units (GPUs) [127, 145]. Another method that we researched was to distribute the EIR workloads via device lending [74, 126]. Both methods improved the processing performance significantly [74, 126].

  We showed the potential of multimedia research in the medical field and showed possible further directions and research topics using the EIR system as an example use case [139].

We contributed to two open source projects: *Lire*, in the field of content-based image retrieval [97], and *OpenVQ*, on video quality [157]. We also released the base algorithm of EIR as an open source project (called Opensea [104]).

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 EIR, we researched several prototypes and techniques to make it easier and more efficient for the medical experts to transfer their knowledge to our system. For this, we explored and developed semi-supervised and cluster-based annotation tools [4, 144]. Based on the findings of one of our annotation tools, we developed a model that can be used to understand events in endoscopic surgery videos better than before and annotate this videos more efficient [49].

- **Contributions to sub-objective 2:** As the basis for the EIR system, we developed a search-based classification algorithm that uses global image features, reaches good classification performance and is very fast at the same time [136]. We developed the theory of intentional framing, which can help to explain why people take pictures and what they want to achieve with them [136]. We researched a method that can be used to accept or discard crowdsourcing workers for content annotation tasks by combining search-based classifiers with crowdsourcing information [141]. We created and researched a prototype of an intent-based video streaming system that uses the intentional framing method to save bandwidth and preserve quality of experience for video streaming [131]. We researched how the search-based classifier can be used to detect and synchronize events in image collections [196, 195]. We researched how the context (a certain watching situation) influences the quality of experience for users when they are watching videos. As a use case, we started with watching videos during a flight. We hosted a MediaEval benchmark task [138] about this topic and published a dataset [137].

  Based on the use cases addressed in the thesis and the EIR system itself, we showed that the search-based classification algorithm is well suited to be applied to several different use cases that involve image classification problems [136, 141, 131, 196, 195, 138, 137, 142, 143, 127, 145, 126].

- **Contributions to sub-objective 3:** We researched different types of visualization for the output of the EIR system. The visualization includes a specific, for research and medical experts developed application [4] and an easier-to-use, web-based version [4, 145]. The visualization approaches can visualize all possible outputs of the EIR system [142].

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

  We researched how multimedia and art can be combined to make people understand disabled people in a better way by developing a game that allows the player to experience a house from a blind person's point-of-view [140].

  We developed a serious game that can simulate the functionality of an eye-tracking device. Based on a crowdsourcing study, we could show that the data obtained by the game

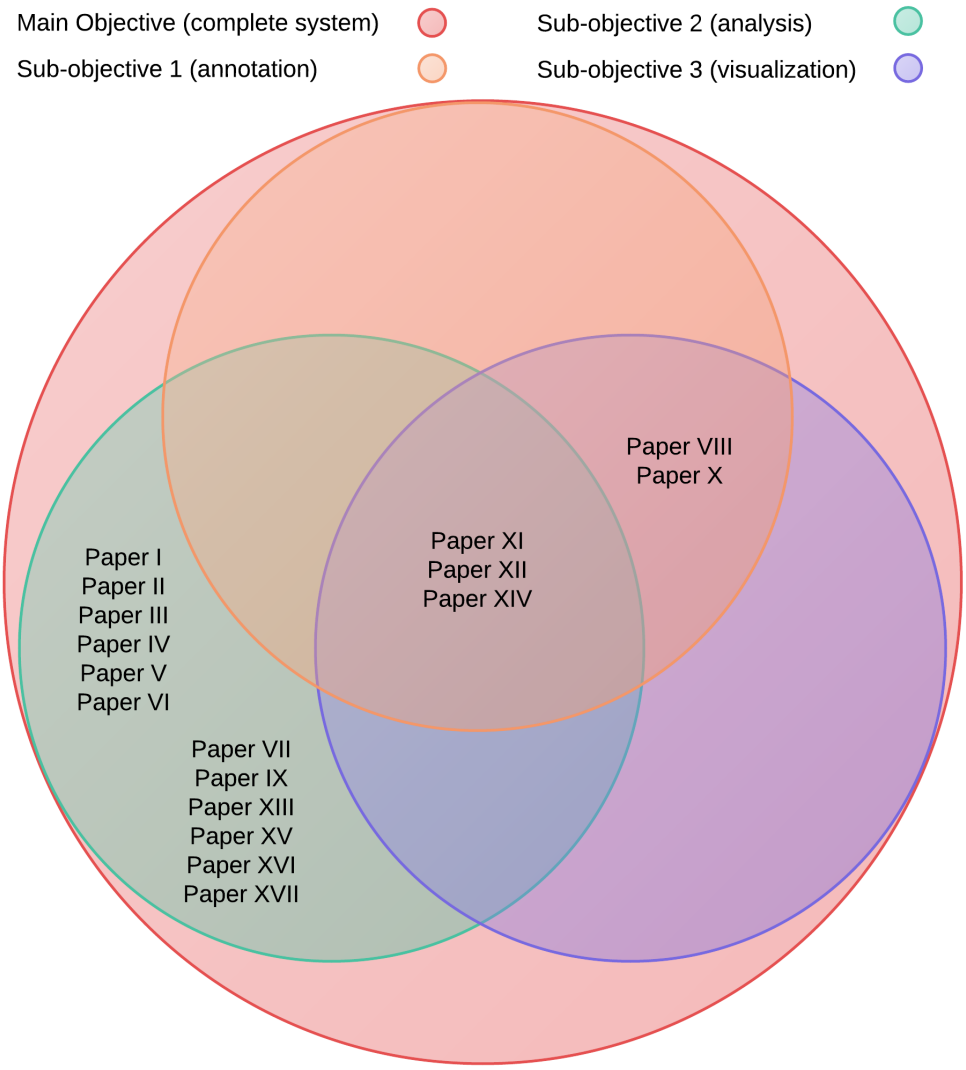



Figure 1.9: This diagram depicts the contributions for each of the in part II attached papers to the, for this thesis defined, objectives.

can be used to find areas of interest in images [134]. We also published the data obtained in this study as a publicly available dataset [133].

We researched how serious games can be used to make scientific content better accessible for the broader population. Therefore, we developed two game prototypes and tested them with real users [98, 108].

We explored how multimedia methods can be used to find manipulations in online content like images and videos, and how to verify that the content has not been manipulated [13]. We also published a dataset that we collected during this study [26].

We researched how the design of complex crowdsourcing tasks can influence their outcome on the use case of 3D reconstruction of soccer players. We published a best practice paper about design and test of such tasks [191] and published a dataset obtained in the study [132].

We looked into the problem of how crowdsourcing can be used in subjective studies such as quality of experience in videos. For this, we looked at different tiling strategies in a

football video streaming system [42]. Based on this research, we also investigated how too much control can have a negative influence on a crowdsourcing study and reported our findings, showing that crowdsourcing can lead to a self-fulfilling prophecy [135].

In addition to the above contributions, the author also supervised several master students, organized workshops and was part of program committees for conferences. We also collaborated with the Cancer registry of Norway in a project that tries to increase awareness for human papillomavirus infection (HPV) and cervical cancer. The cancer registry of Norway started a big user study based on the application, which we helped to develop in December 2015 [108, 98].

## 1.6 Outline

We decided to focus first on workloads in terms of annotation, analysis and visualization. After we developed methods that can be used for our use case and achieve promising results, we wanted to investigate the performance and how it could be improved using parallel processing.

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

**Chapter 2: Medical Multimedia Systems:** We give more background information about the GI tract use case that we chose as a first target for the EIR system. We also present related work focused on other medical multimedia systems and methods. This is followed by a presentation of the basis algorithm of EIR and how we developed it including an overview of the related work and used techniques. Finally, we conclude with a discussion about the papers explaining the search-based algorithm and the use cases that we used to evaluate it in detail.

**Chapter 3: The EIR System:** We describe the complete EIR system. The description is split up in the three main parts. Firstly, we give a general overview, followed by a description of the annotation, detection and localization and visualization subsystems including some experimental results and discussion of real world scenarios for the system. We also describe our improvements in terms of processing performance after the EIR system worked as intended and could be used for testing. Further, we discuss the challenges of real-time distribution, which we faced for this thesis. This is followed by a description of the GPU- and device-lending-based improvements of the performance.

**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 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

In general, a medical multimedia system is an interactive system, that 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 [139]. An important point to add to this definition is that some medical information systems may be fully automatic, but we still consider them interactive, since a medical professional and/or a patient must be in the loop to provide input, interpret and act on the results.

When we started looking into the field of medical multimedia systems, we quickly found out that no complete medical multimedia system for analyzing multimedia data containing information about any parts of the GI tract in real-time exists. A complete system has to fulfill several requirements. These requirements include:

- (i) Efficient real-time processing capabilities, which means that the system should be able to process data so fast that it can be used by medical experts during live examinations. This includes being able to process data produced by standard colonoscopy, which has higher resolutions and needs to be processed in real-time, and being able to process data produced by pill cams (VCEs), which differs in resolution and amount of data to process from the standard colonoscopy but does not have to be processed in real-time.

- (ii) A pipeline for the complete system, which means transferring the medical knowledge into the system using annotation tools, an automatic analysis part and a part for the presentation of the systems' output.

- (iii) Possibility to extend the system with different diseases like bleeding or tumors.

Detection of diseases in the GI tract is mostly focused on polyps. The main reasons are the lack of data related to other GI tract disease indicators and that polyps are predecessors for CRC, which makes them more interesting from a medical point of view since an early removal of a polyp significantly decreases the chance for CRC. Automatic analysis of polyps in colonoscopies has been in focus by research for a long time and several studies have been published [184, 187, 183]. However, not many complete systems exist, and none of them is able to perform 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.

## 2.1 Gastrointestinal Tract Case Study

The GI tract is a complex system and can be affected by various diseases where CRC is one of the most important and a major health issue world wide. Some examples of these diseases have already been depicted in figure 1.2. For this work, we focus on polyp detection since this is the use case with the most available data, but we also give a proof-of-concept for multi-disease detection with a smaller dataset from one of our partner hospitals. CRC is, as mentioned before, one of the most severe disease in the GI tract and often caused by not detected polyps.

### 2.1.1 Colon Polyps

Polyps, as shown in figure 2.1, can be found in different parts of the body like the GI tract, nose, urine bladder or stomach. A colon polyp is a cluster of cells that can occur on the wall of the colon and often sticks out of the wall as a small hill like structure [160].

Colon polyps are often harmless, but over time they can develop into cancer, which is fatal if not detected early enough. Polyps can be developed at any age, but the chance is higher if a person is older than $50$ years [160]. Moreover, a overweight or smoking person has higher chances to get polyps. These polyps do normally not cause any symptoms, and therefore, it is important to participate in regular screening. Early detected polyps can normally be removed completely without any long term problems. Polyps can be separated into three main categories, i.e., *adenomatous*, *serrated* and *inflammatory* [166]:

- A polyp is called *adenomatous* if it is in an early stage of cell change. An example for an *adenomatous* polyp can be seen in figure 2.1. Around two thirds of all polyps are *adenomatous*, but only a small percentage of them develops to cancer (becomes malignant) [91].

- A polyp that has a jagged edge, as shown in figure 2.2, is called *serrated*. Depending on the size and the location of the polyp, it might become cancerous. Smaller *serrated* polyps in the lower part of the colon, also called hyperplastic, are not very often malignant (in a dangerous state). Bigger *serrated* polyps, that are normally located in the upper colon and are often also very flat, have a high chance to be precancerous [91].

- An example for the last type of polyps, the *inflammatory* polyps, can be found in figure 2.3. These polyps are often caused by other diseases like Crohn's disease or ulcerative colitis. *Inflammatory* polyps are not as dangerous as for example large *serrated* polyps, but the disease that are the reason for them can largely increase the overall risk to get CRC [176].

In this thesis, we do not distinguish between the different types of polyps, simply because medical experts will remove every polyp over a certain size that they detect because of the risk of getting cancer at a later stage without knowing in beforehand. Nevertheless, future improvements of EIR could include different classes for different types of polyps to, for example, improve the reporting of findings.

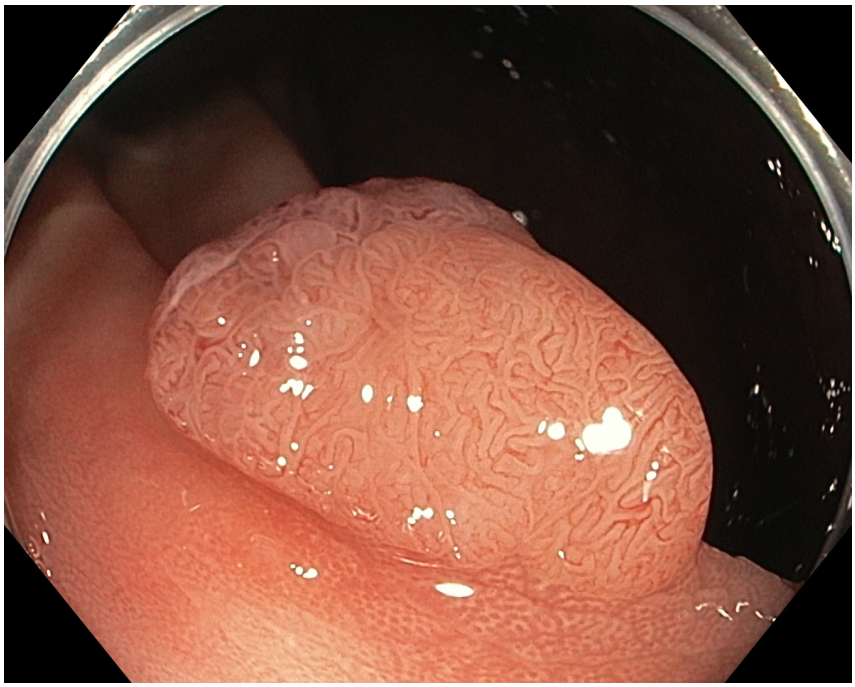


Figure 2.1: An image of an *adenomatous* polyp taken during a colonoscopy.

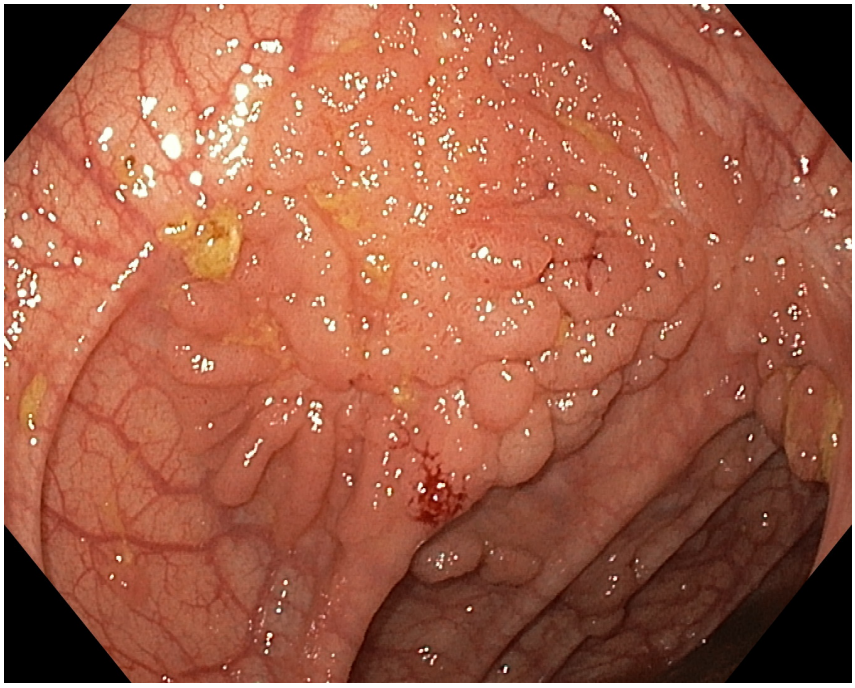


Figure 2.2: An example of a large serrated and flat collection of polyps in the colon.

### 2.1.2 Colorectal Cancer

The development of CRC can be divided into five stages [128]. CRC is in stage 0, if it has not grown beyond the inner layer of the colon, which makes it hard to find in this stage. In stage 1, CRC has grown beyond the mucosa (the membrane on the wall of the colon) but it did not spread to lymph nodes or other places in the colon. In stage 2, the cancer has grown trough the wall of the colon and began to spread to nearby organs and lymph nodes. In stage 3, the CRC

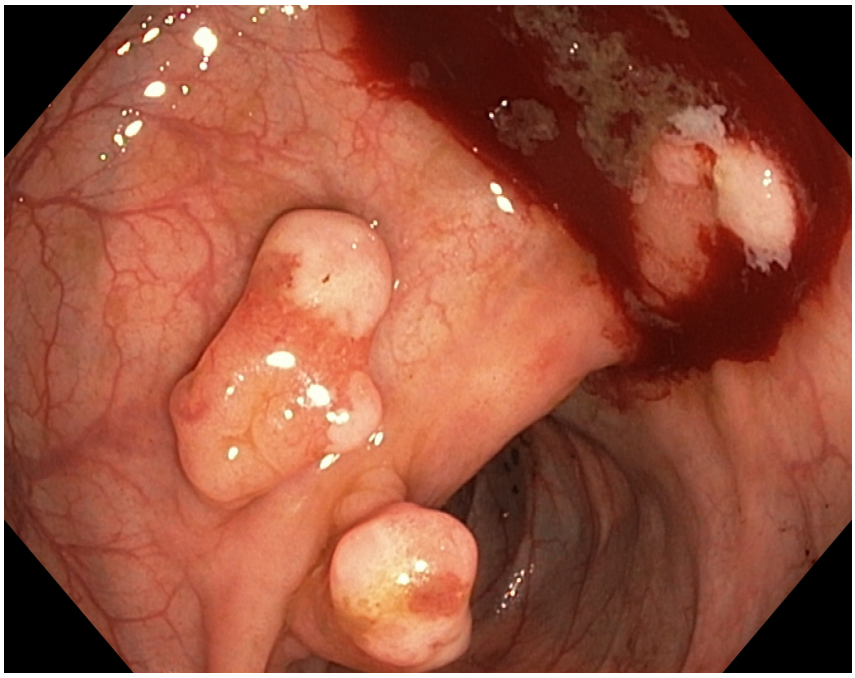


Figure 2.3: A picture taken during a colonoscopy that shows a large inflammatory polyp.

has grown trough all layers of the mucosa and spread close to several lymph nodes and organs but did not start growing on them. In the final stage $4$, CRC has continued to spread and started growing on several lymph nodes and more than one distant organ (such as the lung or liver).

As mentioned before, if CRC is detected at the early stages $0$ or $1$, the prognosis for patient survival is substantial improved. The $5$-year survival rate (the chance of surviving the first five years after the prognosis) for the early stages is $90\%$ compared to $5$-$10\%$ for stage $4$ [128]. Several studies have shown that large population-based screening improves the prognosis and even reduce the incidence of CRC by detecting polyps, which are often precursors of CRC, and CRC in early stages before they can develop into more severe stages [64]. Therefore, the current European Union guidelines recommend screening for CRC for all population older than $50$ years [181]. GI endoscopies are common medical examinations visualizing the lumen (the passage of the colon) and the mucosa of the entire GI tract to diagnose diseases [99].

### 2.1.3 Colonoscopy

GI endoscopies are common medical examinations where the lumen and the mucosa of the entire GI tract are examined to diagnose diseases [99]. 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 with a light bulb (normally around 300W) as shown in figure 2.4. The video signals are transferred to a High Definition (HD) liquid-crystal display (LCD) screen. The most common, gold standard[1] GI endoscopic examinations are gastroscopy (entering via the mouth) and colonoscopy (entering via the anus).

It is important to point out that such endoscopies are demanding invasive procedures, and

[1]The best available standard for an examination of a certain area in the body under reasonable conditions.

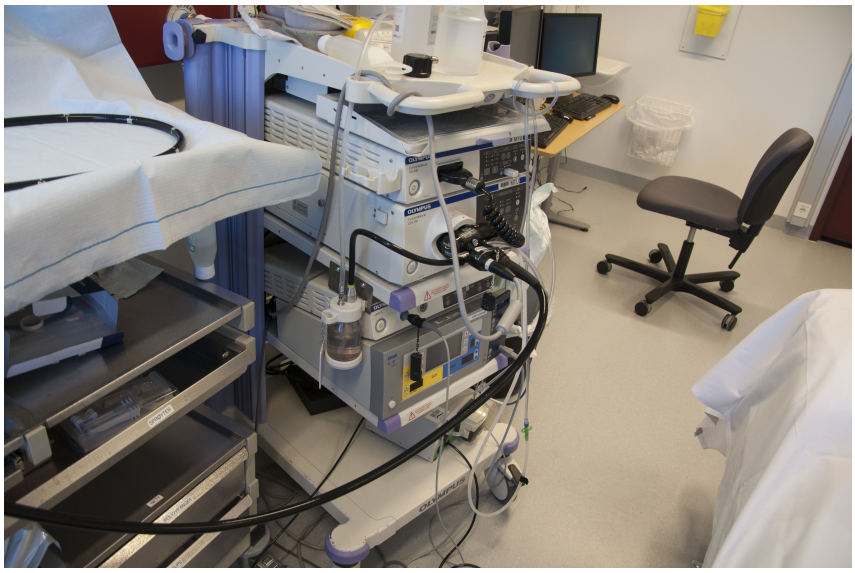


Figure 2.4: An endoscopic processor connected with the colonoscopy endoscope.

can be of great discomfort for patients. They are performed by medical experts, are time-demanding and therefore do not scale well to a larger population. Furthermore, colonoscopy is not the ideal screening test, and in average, 20% of polyps are missed or incompletely removed meaning that the risk of getting CRC largely depend on the endoscopists' (a medical doctor trained for examinations with endoscopes) ability to detect and remove polyps [70]. We therefore aim for a system that detects endoscopic findings in videos of the GI tract where the idea in this context is to assist endoscopists during live examinations.

### 2.1.4 Wireless Video Capsular Endoscopy

As discussed and shown in the sections 1.1, 2.1.1 and 2.1.2, using only standard endoscopy to regularly screen a population is impossible due to high costs, time consumption and lack of high quality medical personnel.

Furthermore, medical screenings used to identify undiagnosed diseases in large populations are debated with known problems like too many false positives, extensive over-diagnosis of diseases that would otherwise clinically not emerge, invasive screening procedures, and high costs [70]. Nevertheless, benefits of screening might outweigh disadvantages, and we will investigate certain screening problems where we conjecture that big data algorithmic solutions might have practical applicability.

A solution that can make it possible to conduct more efficient and large scale screening is using VCEs. A VCE, also often called camera pill, is a small capsule type device (typically $11\text{mm}\times25\text{mm}$), which can have an image sensor, bleeding sensor, pH-sensor, antenna, battery, light source and wireless transceiver. The pill is swallowed (figure 1.7) in order to visualize the GI tract for subsequent diagnosis and detection of GI diseases.

Thus, a person might soon be able to buy a camera pill at the pharmacy (industry estimates a price of $\$10$ for the hardware if mass produced), connect and deliver the in-body video stream from the GI tract to the phone over wireless. The video footage can be pre-processed on the phone to, for example, enhance the quality or do some initial analysis before it is delivered to a

large processing back-end, which does the resource demanding processing steps of the analysis. Therefore, the goal is that the first-order screening analysis is available about eight to twelve hours later (the time the camera typically spends traversing trough the GI tract) possibly without any involvement form medical experts, i.e., a fully automated screening.

### 2.1.5 Medical Data

Nowadays, multimedia data often comes in large amounts (a lot of different data types, easy to produce, etc.). This is also not different for multimedia data produced in the medical field with the additional challenge that most of this data does not contain any information about what it is representing. This leads to the problem of small and very limited annotated datasets in the medical field. Nevertheless, a system that should be able to process medical multimedia data should be created taking large scale data challenges (for example, large workloads, processing time and efficiency, etc.) into account [142].

Large scale data or big data are terms that are commonly used to describe huge datasets that contain so much data that it is difficult for standard database software and operating systems to handle them. The applications are often not able to manage, store and analyze these large datasets in an efficient way or at all. Finding an exact and general definition about when a dataset is big data is hard to find. It often depends on the used technology and the use case. Moreover, because of the fast technological development it normally ranges from a few terabytes to dozens of petabytes [101].

Classification and search in large scale datasets is a timely and resource costly task, but more data leads usually to better classification performance, especially for neural-network-based approaches, which are lately an often researched topic in different fields, for example, in medical imaging, content-based retrieval and social computing. A commonly used definition for large scale datasets is that they are large scale if they cannot fit into the memory of a standard desktop computer [90, 101]. A standard desktop PC nowadays can have easily around $128$GB of memory. Using images as an example and the largest available image size on Flickr as a reference, which is 2048*-pixel for the longest side of the image, results in a file size of between $3$ - $33$MB. This leads to a large scale dataset scenario that would start with around $4,000$ - $110,000$ images, which seems to be a reasonable size [39]. It is important to point out that this definition has to be seen critical because these numbers are, compared to datasets obtained by Google or Facebook, still tiny. Nevertheless, for the medical use case we chose, and for many other scenarios this definition is a reasonable starting point.

It is easy to recognize that annotation, classification and search in such environments are challenging. In large scale datasets, tasks like feature extraction, preprocessing, learning a model or creating annotations can take a lot of time and cost a lot of resources [88, 175, 86]. In the medical field of the GI tract, no large scale datasets are public available, especially, datasets with annotations for our targeted diseases. The reason is that it is very hard to find medical experts who are willing and able to annotate millions of data objects like images. Furthermore, patient related data often comes with data protection challenges. This challenges make the evaluation and comparison of results very difficult.

The only public available dataset containing a large amount of annotated data for our GI tract use case is the ASU Mayo Clinic polyp dataset [168]. We used this dataset as training and

test data for all experiments in this work related to the GI tract use case to make our system as much comparable as possible. This dataset is the biggest publicly available dataset consisting of $20$ videos, converted into a total number of $18,781$ frames with up to $1,920 \times 1,080$ pixels resolution.

### 2.1.6 Filling the Gap

Using the GI tract as a first case study[2], we aim at developing accurate algorithmic diagnostic and intervention technologies that might contribute to increased survival rates and reduce the occurrences of more advanced cases of diseases through standard colonoscopy and capsule endoscopy (camera pill). Our ambitious goal is to develop an end-to-end solution (from learning from the medical experts, over automatic analysis using the search-based classification algorithm, to support them via computer aided diagnosis) where standard colonoscopes and a next generation of camera pills (to be developed by various vendors such as Olympus and Given) transmit high-quality videos of the GI tract that are classified and annotated in real-time.

## 2.2 Medical Image Analysis and Abnormality Detection

As shown in [142], several algorithms, methods and partial systems have been proposed and have achieved at the first glance 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 amount of training and testing data, making it difficult to generalize the methods beyond the specific dataset and test scenarios. Therefore, overfitting (adjustment to random features in the data) for the specific datasets can be a problem and can lead to unreliable results.

### 2.2.1 A Short Overview of Machine Learning

In this thesis, we touched four of the most popular machine learning approaches used for classification of multimedia data. Figure 2.5 gives an overview of them. Support vector machines, instance-based algorithms and clustering are well researched and can be counted in the category of traditional machine learning. Deep learning is a rather new approach that has become very popular lately.

In machine learning, algorithms can be separated into supervised and unsupervised algorithms and two-class and multi-class algorithms. Supervised means that the algorithm needs training data to be able to learn future predictions for data points. Unsupervised algorithms do not need training data, but it is often hard to explain the outcome of the algorithms because the final label or class is not known. Therefore, unsupervised algorithms are often used to explore and understand data without labels [58].

Two-class algorithms can predict if a data point belongs to one of two classes. For example, if an image contains a cat or not. Multi-class classifiers are not limited to two classes and can for example decide if an image shows a cat, dog or bird [109, 40].

[2] As seen by the great disparities of the images in figure 1.2, there will not be one screening filter that can detect all irregularities, meaning that a full system will eventually consist of a large set of pipelined/parallel filters.

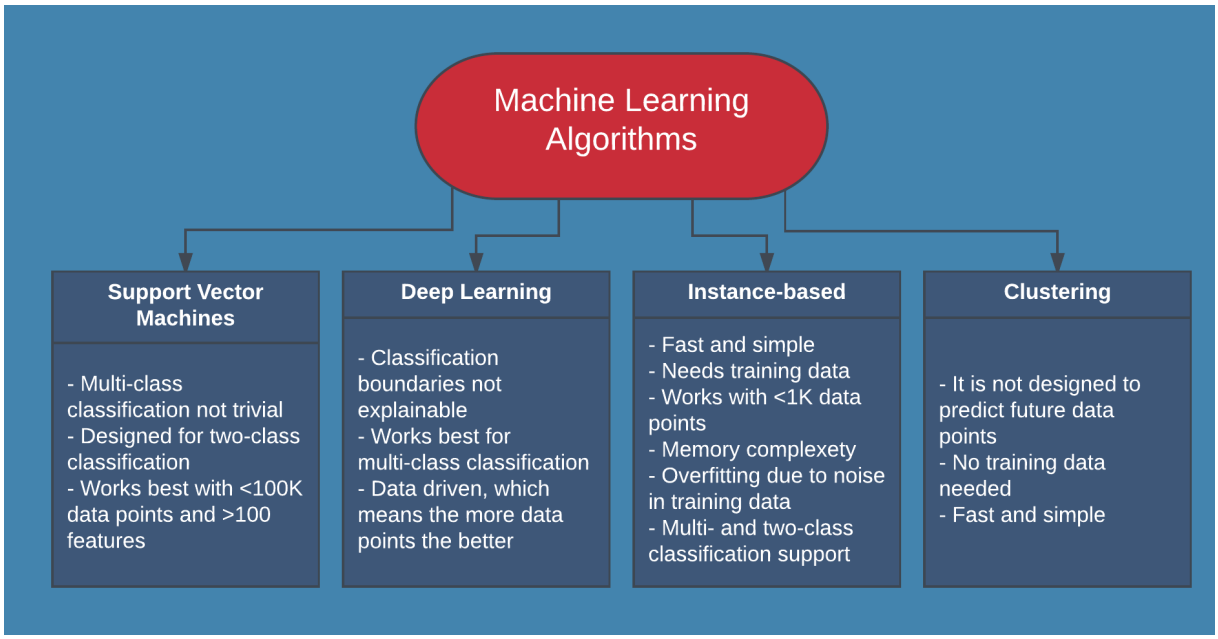



Figure 2.5: An overview of the for this thesis relevant machine learning algorithms including their most important features.

#### 2.2.1.1 Support Vector Machines

Giving a set of training data with two labeled classes, support vector machines (SVMs) try to create a model that can classify future data points with similar features into the corresponding classes. The classification is done by mapping data points as points in a space in a way that the two classes are clearly separated. By design, SVM is not trivial to use for multi-class classification since it is basically made for two-class classification. It can be used for multi-class classification by training one model for each class and combining the results after the classification. SVMs work best with less than 100,000 data points, which makes them useful for small-sized datasets. A downside is that they need a rather high number of dimension for each data point to make accurate predictions [163, 40].

#### 2.2.1.2 Deep Learning

Deep learning algorithms are based on neural network techniques 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, which makes them "deep". 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 [79]. Deep learning can work very well for multi-class classification. Disadvantages are 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 [79, 109, 40]. A more detailed discussion about deep learning in context with the medical area and our use case can be found in section 2.2.3.

#### 2.2.1.3 Instance-based

Instance-based algorithms learn from previous known and labeled training data. This is basically done by comparing new data points with data points from the training data to make a prediction. An often used instance-based algorithm is for example the K-Nearest Neighbor algorithm (K-NN). K-NN tries to find the K nearest data points from the training data for a new data point using a similarity function, e.g., Manhattan distance, to make a prediction. The advantages of instance-based algorithms are that they are fast, simple to implement and that they can achieve good results with even a small amount of training data (depends on the used algorithm). The disadvantages are that they need good training data and that they have a high memory complexity (because they have to remember the attributes of the training data) [109, 40, 2].

#### 2.2.1.4 Clustering

Clustering is a well know machine learning method that is used for unsupervised learning. Clustering algorithms are not designed to predict future data. They are rather used to separate data into different clusters based on attributes and similarity functions. A clustering algorithm could for example separate images into different clusters based on their color features. The number of clusters, used similarity functions and used features depends on the algorithm, data and use case. In general, they are fast, simple to implement and a good tool for exploring data without labels [109, 40].

### 2.2.2 Machine Learning for Automatic Detection of Diseases in the GI Tract

For classifying video endoscopy imaging data, most approaches rely on using an SVM or instance-based two-class classifier in some way. The used image features vary a lot depending on the approach. Some methods use physical dimensions, grayscale intensity values, gradient orientation, RGB color information or texture as input for the classifier. Each of these approaches has advantages and disadvantages. In general, it can be distinguished between two different approaches for the automatic detection of disease in the colon. These approaches are geometrical analysis and machine learning. They could both be used for imaging data that was recorded with a conventional colonoscope or with a VCE. Moreover, it is also possible to use these methods with data from a virtual colonoscopy. Virtual colonoscopy uses x-rays to get 2D and 3D images from the rectum to the lower end of the small intestines. However, these data is significantly different from camera recorded data and therefore out of scope of this work.

Table 2.1 presents a summary of the most relevant approaches in colonoscopies and polyp detection. The last row of the table shows, our system's performance (on the ASU Mayo dataset) to give a comparison. The first approach from Wang et al. [186] is the most recent and best working one in the field of polyp detection, and we co-authored together with them [139]. As one can see in table 2.1, different methods provide different metrics for measuring the performance and use different datasets for training and testing. Moreover, almost all of them focus on polyps.

| Publ./System | Detection Type | Recall / Sensitivity | Precision | Specificity | Accuracy | FPS | Dataset Size |
|---|---|---|---|---|---|---|---|
| Wang et al. [186] ◇ | polyp / edge, texture | 97.70% | N/A | N/A | 95.70% | 10 | 1.8m frames |
| Wang et al. [185] | polyp / shape, color, texture | 81.4% | N/A | N/A | N/A | 0.14 | 1,513 images |
| Mamonov et al. [100] | polyp / shape | 47% | N/A | 90% | N/A | N/A | 18,738 frames |
| Hwang et al. [68] | polyp / shape | 96% | 83% | N/A | N/A | 15 | 8,621 frames |
| Li and Meng [84] | tumor / textural pattern | 88.6% | N/A | 96.2% | 92.4% | N/A | N/A |
| Zhou et al. [198] | polyp / intensity | 75% | N/A | 95.92% | 90.77% | N/A | N/A |
| Alexandre et al. [5] | polyp / color pattern | 93.69% | N/A | 76.89% | N/A | N/A | 35 images |
| Kang et al. [71] | polyp / shape, color | N/A | N/A | N/A | N/A | 1 | N/A |
| Cheng et al. [23] | polyp / texture, color | 86.2% | N/A | N/A | N/A | 0.076 | 74 images |
| Ameling et al. [7] | polyp / texture | AUC=95% | N/A | N/A | N/A | N/A | 1,736 images |
| ***EIR*** † | abnormalities/30 features | 98.50% | 93.88% | 72.49% | 87.70% | 300 | 18,781 frames |

†To test the prototype, we participated in the MICCAI 2015 polyp detection challenge. The challenge had two sub challenges: 1) to detect the exact position of the polyp in a frame; and 2) the latency of the system regarding the first occurrence of the polyp until the moment where the system was able to detect it. For both challenges, we positioned our selves among the top three [142].
◇This approach counts a polyp as fully detected if the polyp has been found once in the whole video or images sequences.

Table 2.1: Performance comparison of polyp detection approaches of state-of-the-art systems. Not all performance measurements are available for all methods. Nevertheless, including every available information gives an idea about each method's performance.

Mamonov et al. [100] presented an algorithm for a two-class classifier to detect polyps in the colon. The method is called binary classification with pre-selection, and it aims at reducing the number of frames that need to be manually inspected. The algorithm processes separate input frames and classifies each frame to either contain a polyp or not. The assumption is that polyps can be generalized as "something that bumps out". This finding was evaluated on a dataset created from frames of videos obtained from five different patients. Based on these experiments, the algorithm reached a sensitivity of $81.25\%$ per polyp at a specificity level 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 all $18,738$ frames that did not contain a polyp was $9.8\%$. Assuming that it is normal to have multiple frames available for a single polyp, these numbers seem quite promising [142].

A similar approach is presented by Hwang et al. [68]. 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. Ellipses are then fitted into the segments by constructing a binary edge map for each segmented region and using a least square fitting method. A threshold-function is used for the creation of the edge map. Regions with too little edge information in their respective edge maps are discarded. These ellipses are then further evaluated for matching of curve direction, curvature, edge distance and intensity [142]. The direction of the parabola from any part of the ellipse must be matching the direction of the corresponding part of edges for the ellipse to be considered a polyp. This assures that the detected edges build an ellipse-like shape instead of, for example, a parallel one. Furthermore, the curvature of the ellipse is split into six parts. At least two adjacent parts must have a strong edge pattern, otherwise, the ellipse is discarded [142]. 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. This allows to apply a threshold in number of frames for the detection to reduce the number of false positives. 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 [100], 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. It is important to mention that this assumption depends largely on the frame rate of the camera that is used for recording the video [142].

The most recent and complete system in the well researched polyp detection field is Polyp-Alert [186], which is able to give near real-time feedback during colonoscopies. This approach is also listed as number one in table 2.1. 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. A polyp is counted as detected correct if it has been found once in a frame in the whole sequence of the video, which makes it easier to achieve a high sensitivity compared to frame wise evaluation (each frame containing a polyp is used for the evaluation). Unfortunately, the dataset is not publicly available, and therefore, a detection performance comparison is not possible. Compared to our system, this system seems to reach higher detection accuracy, but our system is faster and can detect polyps in real-time (see table 2.1 for details). 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. It is also designed to support the goals of the project in terms of scalability, flexibility, due to the support of VCE and standard colonoscopy, and as a massive scale screening option as a preventive service.

Other papers that discuss how to improve performance of endoscopic surgeries in general (not colonoscopy) are for example [112, 110, 111]. 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, making them interesting for our system for future improvements.

### 2.2.3 A New Trend - Deep Learning

Since deep-learning-based approaches are commonly used nowadays, they are also discussed in relation to the GI tract analysis. Deep learning in the field of medical imaging is a very popular topic that emerged recently, and many researchers believe that it holds a lot of opportunities to improve medical imaging [47]. The basic ideas of deep learning methods are conceptually easy to understand (but hard to master), and lately, a large amount of academic research has been performed in this direction. Results recently reported on, for example, the ImageNet dataset look quite promising. It seems that deep-learning-based approaches outperform other machine learning methods. This is true for some of the image annotation and object detection problems [29].

Nevertheless, deep learning comes with some challenges that make it not straight forward to use for the GI tract use case [24]. Firstly, training is very complicated, needs a lot of training data and can take a long time. Our system has to be fast and understandable since we deal with

patient data and the outcome can decide between life and death [142]. Therefore, a *blackbox*[3] approach, which deep learning approaches can be, seems to be the second best way to solve a problem that has to be understood very well by all involved users. Not being able to explain how a certain decision has been made can lead to serious problems in the medical field since it is not possible to evaluate the output and the decision made by deep-learning-based systems properly yet and there will always be a chance that a deep learning approach will completely fail without being aware of it [113]. The best way is still to understand the problem and then solve it [142].

Further, they require a lot of training data. In the medical field, this is a very important issue since it is hard to get ground truth 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 [148]. Finally, deep-learning-based approaches 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 (class boundary) is an important information [142]. 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 [162].

### 2.2.4 Current Limitations in Medical Multimedia System

In summary, all the related work shows promising results with a lot of different approaches for our use case of disease classification in the GI tract. Most of the approaches use machine learning and are very focused on the image processing part for medical imaging. Moreover, deep learning approaches seem to be the new area of interest and in the near future most probably a lot of research in this direction will be conducted. Nevertheless, both directions come with challenges. They are either; (i) too narrow for a flexible, multi-disease detection system; (ii) have been tested on a too limited dataset not showing if the methods would work in a real scenario; (iii) need a large amount of training data that does not exist or is hard to get like the deep learning approaches; and finally (iv) provide a too low performance for a real-time system or they have ignored the processing performance aspect in their evaluations at all. In the following section, we will present a promising approach that addresses all the shortcomings of the actual existing methods by using a simple but accurate, fast and easy to understand search-based classification approach embedded in a complete medical multimedia system. Furthermore, as we demonstrated with our evaluation, the search-based approach does not require a large amount of training data as for example deep learning approaches and can, because of its architecture, easily be extended with different types of data to analyze (sensor data, patient records) and diseases to detect.

[3] Using a system or a method without really knowing or understanding what exactly happens between the input of data and the output of the results.

## 2.3 The Basis of Our System: A Search-based Classification Approach

As discussed previously, to visually detect polyps and other diseases in the colon is a very demanding and exhausting task for the doctors. Often, they have to perform several colonoscopies during a shift, and at the end of the day, they are tired and can naturally not concentrate in the same way as in the morning. With EIR, we developed a system based on content-based information retrieval methods that is able to support the medical doctors as a *third eye* and improve their performance, and with that increase patient care and surviving rate.

Such a system has to be very fast because it has to be able to assist the medical experts during live colonoscopy examinations. Furthermore, it should be easy to operate by the medical experts, be able to batch process a large amount of data for the VCE data and its output has to be as accurate as possible.

To fulfill these requirements we developed a novel and search-based classification approach, which serves as the basis of the EIR system. Among all other possible alternatives, search-based classification is promising because it is easy to understand, efficient in terms of processing time, easy expandable with new diseases and at the same time reaches very food classification accuracy. Apart from that it can easily be extended by other data types such as sensor data, patient records, etc. that can be classified using the same retrieval methods. The search-based classifier utilizes global image features (a mathematical, intermediate representation of the image content based on different attributes, for example, color distribution or texture attributes, examples can be seen in figure 2.6). We decided to use global image features, because they are very light weighted, easy to compute and quite easy to understand by humans.

### 2.3.1 Global Image Features

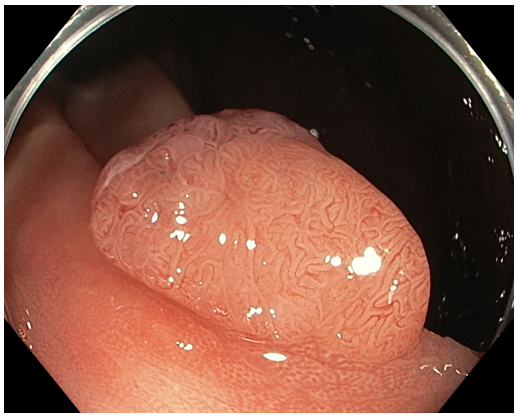
As mentioned before, global image features are features that can capture the overall content of an image in one single feature to describe it, e.g., the color distribution or texture features [95]. In figure 2.6 an original image of a polyp (figure 2.6(a)) is shown with four different representations of how the same image could look using a certain, by a global feature captured, attribute to represent it.

Figure 2.6(b) shows the polyp image how it would look using a global feature that captures the edges of an image. Figure 2.6(c) shows the same picture as a possible color feature representation. In this case, the most prominent colors in a certain area are merged, which leads to a kind of superpixel representation (a superpixel is a more abstract representation of a set of pixels of an image). How a feature that captures the color and the edge information could represent an image can be seen in figure 2.6(d). The last figure 2.6(b) depicts a feature representation of a global feature that captured the texture information of an image.

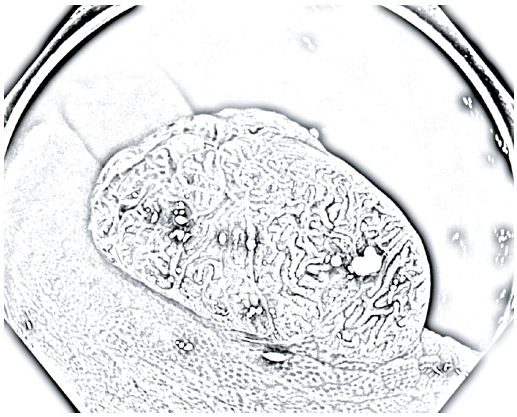
Global features in classification tasks are in some way underestimated because they are very simple, but as shown in for example [136], this can also be an advantage.

Table 2.2 gives an overview of all in EIR supported and tested features. The majority of papers about classification or search for content-based retrieval use mostly local features or features created by neural networks.

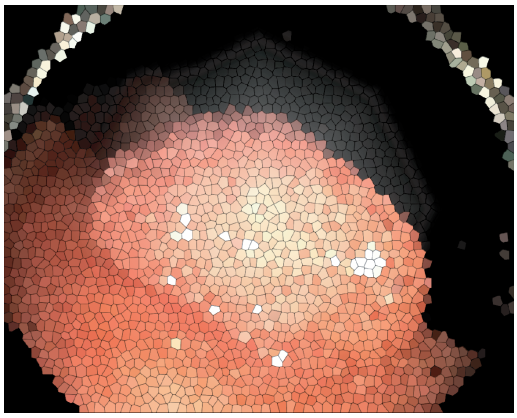
Examples are the well known SIFT (Scale-invariant feature transform) [93], SURF (speeded



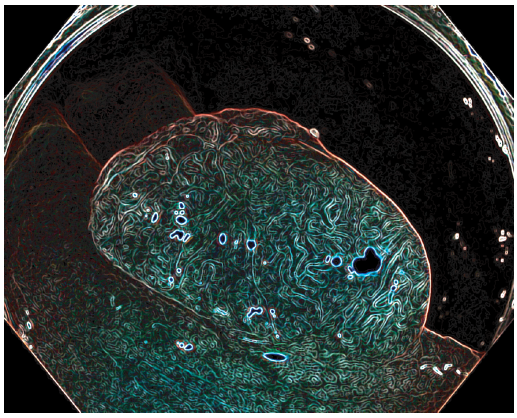
(a) Original image of a polyp.



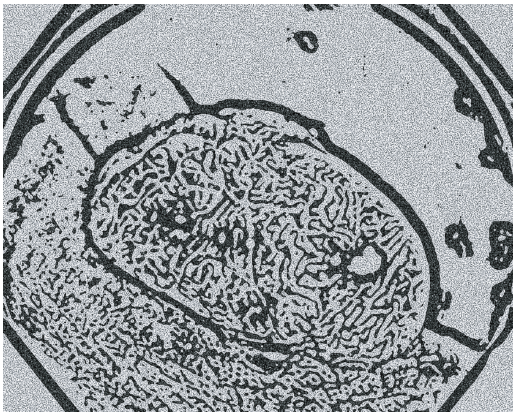
(b) Image of a polyp as an edge-based feature could show it.



(c) Image of a polyp as a color feature could show it (most prominent colors in a region are merged).



(d) Image of a polyp how color and edge feature could represent it.



(e) Image of a polyp how an texture feature would represent it.

Figure 2.6: Examples of a polyp image represented as different global feature representations. It is important to point out that this is not how the features actually look because they are histograms and not meant to be shown as images, but it can help people to get an idea about how they work.

up robust features) [9], FREAK (fast retina keypoint) [3] and BRISK (binary robust invariant scalable keypoints) [83]. Local features are normally used with the bag of visual words approach which uses a text retrieval approach to match objects in videos and images [156].

| Feature | Dimensions/bins | Captures |
|---|---|---|
| JointHistogram | 576 | c, cd, t, jh |
| JpegCoefficientHistogram | 192 | c |
| Tamura | 18 | t |
| FuzzyOpponentHistogram | 576 | fc |
| SimpleColorHistogram | 64 | c |
| FuzzyColorHistogram | 125 | fc |
| RotationInvariantLlBP | 36 | t |
| FCTH | 192 | fc, t, jh |
| LocalBinaryPatternsAndOpponent | 288 | c, t, jh |
| PHOG | 630 | t |
| RankAndOpponent | 576 | t |
| ColorLayout | 33 | c |
| CEDD | 144 | fc, t, jh |
| Gabor | 60 | t |
| OpponentHistogram | 64 | c |
| EdgeHistogram | 80 | t |
| ScalableColor | 64 | c |
| JCD | 168 | fc, t, jh |

Table 2.2: Table of all global features tested and supported by EIR. *Feature* is the name of the feature. *Dimension/bins* shows the size of the feature vector and *Captures* indicates which type of characteristic of the image/frame is captured and incorporated in the feature: *c:* standard color information; *cd:* how color pixels are distributed to each other; *fc:* a fuzzy color scheme; *t:* texture attributes such as edges, gradients or other texture characteristics; *jh:* combine different attributes likes for example texture and color of pixels.

Local features are very successful in detecting objects in an image or videos, and promising results have been reported [147]. What local features cannot really do, is to capture global characteristics of an image or video frame [6, 14]. Thus, for large scale classification and retrieval tasks it is challenging to use local features, because they are very complex, have higher computation time and need more space on the hard disk and the memory [48].

The amount of actual research based on the use of global features for classification is limited. The most relevant related work is done by Oliva and Torralba [117, 118], who describe the role of global features to build the overall meaning of an image. They present a formal approach where global features are utilized to detect the global meaning of an image. They used an encoded, coarse representation of the organization of high and low spatial frequencies computed from the images. These computed features are then used to find meaning of an image by using a spatial envelope model. Based on the work from Oliva and Torralba, two other related works were developed by Hays and Efros. Firstly, they used the original method of Oliva and Torralba to solve the task to estimate the geographic location of a single image [61]. To achieve this, they combined seven different global features, including color, geometry and textons (texture features) to achieve high precision for their system. The overall assumption was that the features they used were usable for geographical location estimation based on the global view on the image gained by them. Secondly, they used their combined global features to find the missing

parts of incomplete images of geographical locations in [60]. To tackle this problem, they performed a search for the missing parts based on the global features, and filled them with parts of the best matching image from the obtained results. From these basic works, a lot of other features have been suggested to identify different characteristics of an image or video frame. In our work, we have considered the following global features (table 2.2):

Joint histogram: The joint histogram is an improved version of the standard color histogram [165]. The joint histogram is created by constructing a multi-dimensional histogram based on a selection of local pixels in the image. The advantage of the joint-histogram-based features is that they are small and easy to compute which makes them ideal for a time sensitive task [123].

Jpeg coefficient histogram: The Jpeg coefficient histogram feature is implemented following the Jpeg standard of the world wide web consortium (W3C). It uses the frequency distribution of the $64$ discrete cosine transform coefficients in a Jpeg file to describe the content of an image. This is performed by counting for all $64$ coefficients in an $8 \times 8$ block the frequency of each value and store these values in a histogram [182].

Tamura: The Tamura feature [169] is based on the assumption that textural features correspond to the perception of the human eyes. Tamura compared coarseness, contrast, directionality, line-likeness, regularity and roughness, which are six different texture features, with psychological measures taken from human experiment participants. The three features that achieved the best results in his evaluation are coarseness, contrast and orientation. Coarseness measures the size texture primitives (also called texture elements or texels) [57]. Larger textures have larger primitives and fine textures have smaller ones. The contrast measures how distinctive the differences between the textures in the images are. The contrast can be considered as clear if all areas can easily be distinguished from each other. The orientation describes the dominant orientation of the textures in the image. A single image can have only one dominant orientation or several of them. Moreover, an image can also have no orientation at all, which then is called isotropic. For the Tamura global image feature, coarseness, contrast and orientation are extracted from an image and stored in a histogram representation [65, 95].

Fuzzy opponent histogram: The fuzzy opponent histogram is a simple and fuzzy $64$ bin opponent histogram, based on the fuzzy opponent color space. The feature is rather large for a global feature but can achieve good results with noisy images [178].

Simple color histogram: The simple color histogram feature provides a simple to compute and based on the standard color space histogram. The number of dimensions (number of values, also sometimes called bins, in the vector describing the feature) is configurable (scalable) and the histogram is normalized to eight bits per bin [95].

Fuzzy color histogram: The fuzzy color histogram is an improved version of standard color histograms. The histogram is computed by a fuzzy-set member ship function that compares color similarity of all pixels with each histogram bin. The advantage is that it has less dimensions than the standard color histogram and it is less sensitive to noise interferences such as illumination [56].

Rotation invariant local binary pattern: The rotation invariant local binary pattern feature is a simple texture-based feature. It applies a threshold on neighboring pixels with a binary number as outcome. These binary descriptors of the texture are the dimension in the feature. The rotation invariant version of the feature is robust against rotations of the images. Because the feature is very easy to compute, it is well suited for real-time applications [116].

Fuzzy color and texture histogram (FCTH): FCTH is a low level feature, which combines the color and textural information of an image in one histogram. Due to its limited size of $72$ bytes per image, it is usable for large scale image databases [22].

Local binary patterns and opponent histogram: The local binary patterns and opponent histogram feature is a simple combination of the rotation invariant local binary pattern and opponent color features. The combination results in a joint histogram [95].

Pyramid histogram of oriented gradients (PHOG): PHOG computes a fuzzy histogram of gradient directions, which is performed in three steps. Firstly, an edge detector is used to detect all edges of an image. Secondly, the algorithm follows these edges and computes the gradient directions for the histogram. This is performed top down following a pyramid pattern. The image is split up into a quad-tree where each part of the tree has four children nodes. For each node of the tree, a histogram is calculated. Finally, all these sub-histograms are concatenated into one large histogram [15].

Ranked and opponent: The rank and opponent histogram is an implementation of a joint opponent histogram combining a $64$-bin RGB color space histogram and a pixel rank [95].

Color layout: Color layout is based on the MPEG-7 color layout, which is a low level feature. It represents the spatial distribution of colors in an image. The main functionality of the feature is to capture the spatial information of the most representative colors of an image and superimpose this information on a grid that is laying on top of the image [19].

Color and edge directivity descriptor (CEDD): CEDD is a low level feature, which combines the color and the edge information of an image into one histogram. The size of the feature is limited to $54$ bytes, which makes it very useful and fast for large scale use cases [21].

Gabor: The Gabor feature is based on the Gabor filter [189], which is a linear filter used for edge detection. The filter represents the image in a to the human-visual-system similar way. This is achieved by the presentation of the frequency's and the orientations contained in an image. The dimensions of the feature are a set of different frequencies and orientations combined in one single histogram [37].

Opponent histogram: The opponent histogram is a $64$-bin opponent color histogram described in Sande et al. [178]. It is a combination of three 1-dimensional histograms, where each of the histograms represent one channel of the opponent color space, which is an alternative representation of colors. It consists of three different parts, the luminance component, the red-green channel and the blue-yellow channel. For the features, all three parts are computed for an image and stored in a histogram [95].

Edge histogram: The edge histogram feature is based on the description in the MPEG-7 visual standard for content description [19, 154]. It provides a description for textures in an image, which are not homogeneous. The feature is constructed in the same way as the color layout feature, but instead of the spatial distribution of the colors, the spatial distribution of the edges in an image is captured. The feature is very compact, scale invariant and can be used for rotation-sensitive and rotation-invariant matching [154].

Scalable color: The scalable color feature provides a description of the color distribution in an image. The feature is based on the MPEG-7 standard and is a color histogram that uses the HSV (hue, saturation, and value) color space. The HSV color space represents color, as the name indicates, by hue, saturation and value, where value stands for the brightness. The representation is quantized to $255$ bins and saved in a histogram [154].

Joint composite descriptor (JCD): JCD is a joint descriptor, which combines two compact composite descriptors (CCD) in one. A CCD is a very compact global image descriptor with the purpose to combine many different features in one descriptor. The for JCD used CCDs are the fuzzy color and texture histogram and the color and edge directivity descriptor. The combination of the two descriptors is possible because their color information originates from the same fuzzy color system. The result of the combination is a descriptor, which contains fuzzy color information, texture information and edge information [20].

To make use of the global features extracted from the images, they have to be used in a certain classifier or retrieval system. In this work, we combine both. Before we can use the global features for our search-based classification, we have to index them and make them searchable.

### 2.3.2 Indexing

From the intermediate format for each global feature (or any other data data type), indexes can be generated (one index per feature or several features in one index). 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 used features. 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 approach [142].

The hashing approach is based on locality sensitive hashing (LSH). The main idea is to use multiple random hash functions to hash the values of the features. Similar images will then get the same hash values and therefore be hashed into the same bucket. 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 [159].

The in this work used hash function $h(v) \in \{0, 1\}$ for a histogram $v$ is defined as $h(v) = sgn(v \cdot r)$, whereas $sgn$ is the sign or signum function (extracts the sign of a real number) and

$r$ is a random vector with evenly 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 $H_j(v)$, $0 \leq j < m$ are generated [142].

The parameters for the hashing-based approximate indexing are chosen based on evaluations on a image dataset consisting of $100,000$ 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 (the search time could be reduced to around $30$ms for searching one image in an index consisting of $1.5$ million images) and at the same time, to a good trade-off between search time, and precision.

### 2.3.3 Search

The standard search for an image that we use in our search-based algorithm is performed on the fly on former created indexes and returns a ranked list as result. This means, for each image a term-based query from the hashed feature values of the query image is created at run-time. Based on these values a comparison with all images in the index is performed. The end result is a ranked list of similar images.

The ranked list is created by a distance or dissimilarity function associated with the low level features. This is done by the computation of the distance from the query image to all images in the index. The used distance function for the ranking is the *Tanimoto* distance [170] 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 position [170]. The final ranked list is used for showing the results of the search or for further tasks. In our case this is the classification step. To be able to make a decision in the classification step, two important aspects have to be considered first. These aspects are, which features to use and how to combine them.

### 2.3.4 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 before the classification step, which features we want to use for a specific use case. This is important, because global features and combinations of them that are chosen in a random way can lead to negative results for the classification or search task. Badly chosen feature combinations can introduce noise (if to 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).

It is important to explore, which features are for the respective task more useful than others. A lot of work has been performed in the field of feature selection, and different machine learning techniques were utilized for it [106]. For example, principal component analysis (PCA)

which is a reduction of a large set of variables to a smaller, conceptually more coherent set of variables that contain the same information as the larger representation [32]. Another example is information gain attribute evaluation (IG), which computes the information gain of a given feature with respect to the classification problem to determine which feature gives the most information [28]. A final example is the SVM attribute evaluation, which ranks the variables of the features using a weight assigned from a support vector machine [52].

Guldogan and Gabbouj [50] tried to reduce the complexity of a system by reducing the number of features. They utilized standard feature selection algorithms, like PCA and IG, to get measures, which show them how good or bad a feature performs for a given task. Based on these measurements they applied majority voting. The output is a ranked list of features that they use to select the final features. Their evaluation results demonstrate that this method can improve the classification performance and at the same time reduce the computation time. In this thesis, we perform a simple feature selection by testing different combinations of features on smaller reference datasets to find the best combinations im terms of processing speed and classification accuracy.

### 2.3.5 Feature Combination

As one can see in table 2.2, EIR supports a lot of different global features and the combination of them. In different use cases, different combinations of the supported features would certainly lead to more accurate results than others. The combination of different features, containing feature spaces with different sizes of dimensions, is a non-trivial problem that has been topic to research for several years.

A sophisticated combination of features can help a system to obtain more accurate classification and search results. Nevertheless, feature combination comes also with some pitfalls that must be taken into account when applied. If feature combination is not performed correctly, it can lead to a decrease of performance [27, 51, 193]. Features can be combined in two different ways. The first one is called feature 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 or *late fusion*. For *late fusion* features are combined after a decision making step.

#### 2.3.5.1 Early fusion

A detailed overview of the pipeline for *early fusion* can be found in figure 2.7. The underlying concept of *early fusion* is to combine uni-modal features, after they are extracted, into a multi-modal representation, which basically means that the feature values will be combined into one large vector for representation. These large feature vectors can be used for search or classification tasks such as supervised learning (labeled train and test data set for learning), two-class classification (a binary decision between two classes) or unsupervised learning (not having any training examples) [106]. The problem with this method is, that this vector can contain a lot of noise that is introduced by not meaningful features [27].

Another early fusion method is to concatenate all features into one long feature vector, but applying feature selection or feature transformation methods, like the in section 2.3.4 described

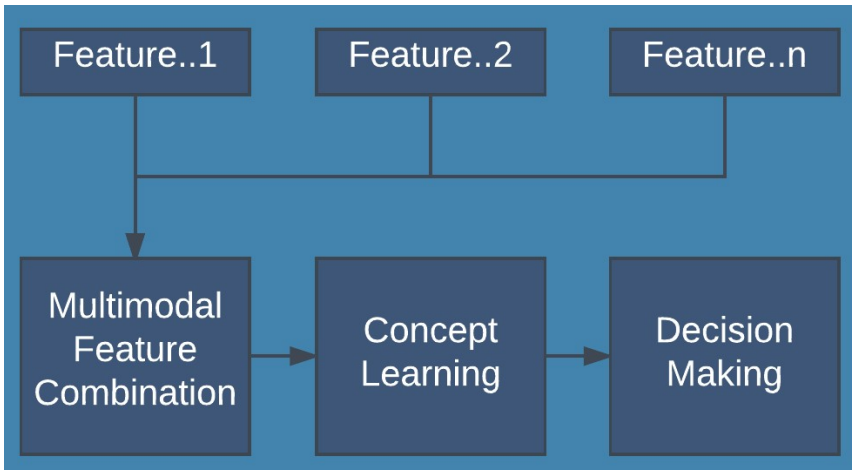



Figure 2.7: Pipeline for *early fusion* of features. The features are first combined into one large vector and then a decision is made based on this fused feature vector.

PCA [32], independent component analysis (ICA) [80], linear discriminant analysis (LDA) [53] or nonlinear component analysis (NLCA) [149] on this vector. Applying one of these methods can lead to a reduced length of the vectors. Machine learning applied on these smaller vectors can be performed in a more efficient and precise way, which can lead to faster processing times and more accurate results. Nevertheless, also with these methods problems can emerge. In particular, the vector can contain noise or some features can be missed that by itself have a bad performance but combined with other features would perform very well or vice versa [193].

Which features should be combined is open to the user and the task that should be solved. For example, one can combine visual, audio, text features and their different variations. A challenge, that has to be taken into consideration, is that a fusion of many features before classification can introduce noise in the data. This happens due to the fact that the more information is combined in a huge vector the more of this information can be meaningless for the classification problem to solve. Another problem is that the combination of very diverse features, like audio with visual features, can be challenging. This challenge often occurs if one tries to combine features that have different dimensions and range of values in these dimensions. For example, combining an audio feature with many hundred dimensions with a visual feature with many thousand dimensions. To tackle these challenges and to reduce the loss information within the fusion process, it is recommended to perform feature selection, feature reduction and normalization methods on the data before the fusion [158, 51].

#### 2.3.5.2 Late fusion

In *late fusion* or also called decision fusion, each feature is processed by an own classifier. After these first classification steps, the output of each classifier is combined to obtain a final result. An overview of the basic steps performed for late fusion are depicted in figure 2.8.

Because each feature is processed in a separate classifier, *late fusion* is very costly in terms of learning effort. Moreover, to combine the pre-classified features one or more additional classifiers are needed. Another challenge is the possible loss of information that comes naturally if different features are combined [158]. The combination of the output of the pre-classifiers is a very important step and can be performed in different ways. Which method is the best depends

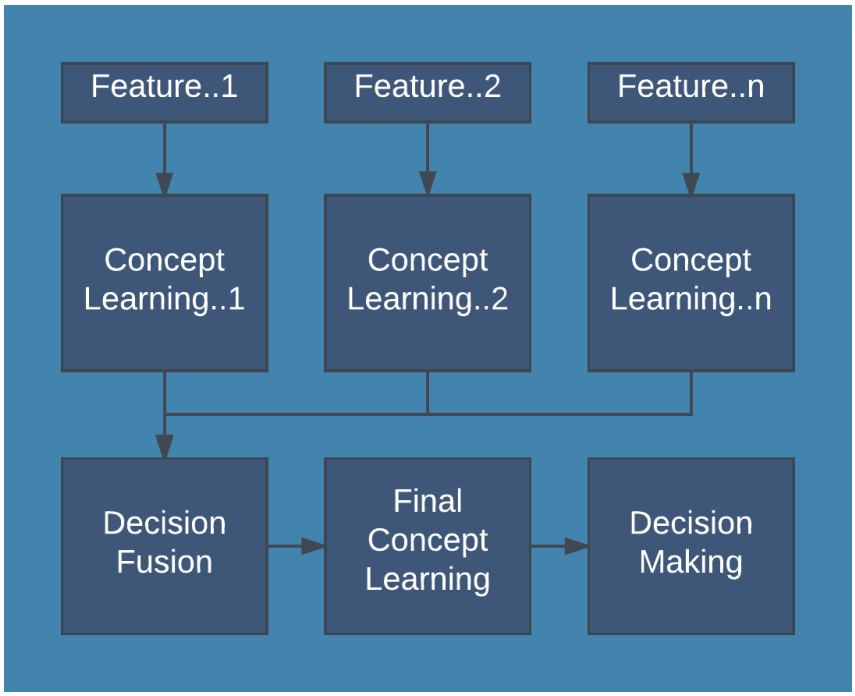



Figure 2.8: Pipeline for late fusion of features. The features are first processed by separate decision making methods (e.g., a classifier) and then combined.

on the dataset, the features that are used and the metrics that are used to calculate the distances between the different features. A sophisticated and well chosen combination strategy can lead to an improvement of the classification results [43]. Trotman [177], Hsu and Taksa [66] and McDonald and Smeaton [103] came to the same conclusion: Some datasets are better suited for *late fusion* based on rank score, for other datasets fusion based on rank or weighted rank is the better choice. Their experiments revealed that a combination using a score metric leads to better results for features that have the same metrics for their distance scores. For instance, if all the features that should be combined use cosine similarity, then it is recommended to combine them by score. On the other side, if they use different distance metrics for the scores, a combination by rank can lead to more precise results.

In [87], the authors try to learn tag relevance for social image retrieval by using multiple features in an unsupervised learning environment. They state that visual content cannot be discovered by using only one visual feature. To prove this claim, they tested different approaches of feature combination methods to find the best one for their task. The first one is based on Borda count, which reduces the scores for each rank by one so that the last item in the rank will get zero points [179]. The second one is called *UniformTagger*, which is a combination of different base learners that are combined in an uniform way. The conclusion is that learning using a combination of different features can outperform single-feature-based learning systems. The second observation is that *late fusion* approaches lead to a better performance than *early fusion* approaches [87, 158]. Escalante et al. [34], who came to the same conclusion, showed in their paper that *late fusion* performs well for multimedia retrieval tasks. They fused multiple, heterogeneous and for annotated collections developed image retrieval techniques. 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 decided to

use it for feature combination in our work. Therefore, we classify each feature that we use separately, and combine them afterwards using a majority decision weighted by the ranked score (an image class in a higher position in the ranked list gets a higher weight).

### 2.3.6 Search-based Classification

Classification is an important topic in the field of machine learning and it is normally implemented trough well known classifiers like Naive Bayes classifiers, SVMs, neural network approaches or binary classifiers [17, 105, 106]. Machine learning has not been used very frequently in combination with retrieval systems, and almost not existent is machine learning based on retrieval methods.

An example for machine learning used to improve content-based retrieval can be found in the paper of Goesselin and Cord [46]. They use Bayes classifiers, K-NN and SVM in combination with an active learning system. The learning system is based on users evaluation of the search results. The users feedback is used in a method called *RETIN AL System Framework*. The framework improves the retrieval results on the fly with the help of the users feedback. Which means that while the users interact with the retrieval system they can steer the system towards more relevant search results.

Related work, which utilizes retrieval methods for classification to improve the performance and accuracy of a classifier is rare. Nevertheless, there exists one related paper from Larson et al. [77]. In this paper, the authors give an overview over three tasks performed at the *MediaEval 2012 Benchmark*. These tasks are all related to automatic tagging and geographical tagging. All three tasks focus on tagging of videos using as much information as possible. The types of information used in the tasks are videos, user-generated meta data, speech recognition, transcripts of audio and images. The tasks are separated into two parts. The first part focuses on the extraction of different features from the data. The second part focuses on assigning labels to data object. This is performed by an information retrieval approach where the label of class is considered as the query for the search performed by the system. The search in this adopted system results in a collection of documents. These documents are then annotated with the same tag as the search query tag. One advantage, which they found in their paper, is that it can be used for supervised (training data is available) or unsupervised (no training data is available) machine learning.

Li et al. [86] tried a similar approach to automatically annotate images. They used a search-based technique without having a training dataset for their system learn from. To obtain annotations for the images textual meta data of pictures, which is stored in a cluster formed by the search results of the query image, is used. Based on these clusters the tag for the query image is assigned. On problem with this method is that the performance of the system seems to be not good and it is not described very well. Moreover, it is not developed for a classification but rather for annotation tasks.

The search-based algorithm developed in this work has been implemented using the Lire [96] open source library for content-based image retrieval, which allowed us to experiment with a large set of global features. Since Lire is based on the Lucene indexes, it also allowed us to create an algorithm that is able to include any type of multimedia data if needed [171]. Lucene inverted indexes are created using k-way merge. 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 reasonable fast to search [171, 150, 73]. The indexes are field-based and the number of fields is variable depending on the number of used features and the fields are stored using LSH as described in section 2.3.2.

The basic algorithm is described in detail in [136] and [142]. In its basis, it is a simple K-NN algorithm and defined as following: Classified as class $c$ is the class with the highest weighted $ClassScore$ of all classes $c \in C$.

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

$ClassScore$ 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 $RankScore$ in the rank list [142]. The here presented 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 or videos to meta data. The main advantages are that it is very simple and easy to understand, achieves state-of-the-art classification results and that it is very fast in terms of processing time, which is demonstrated by applying it to different use cases described in more detail in the following section.

### 2.3.7 Use Cases and Implementations

The first use cases that we addressed was the area of user intent and the human perception of multimedia content. This area has been chosen because of two main reasons. Firstly, it requires processing of a huge amount of data and these data is easily available due to Flickr, Youtube and Twitter. Secondly, it was the first evaluation scenario for our search-based classification method.

At the beginning, we developed the theory of intentional framing [136], and a first implementation of the search-based classification method tested on huge datasets (1.5 million images). We also applied our method on other scenarios such as social event detection and refinement of crowdsourcing votes [195, 141, 92, 196]. We also developed a multimedia system which uses the intent information of online videos to save bandwidth by reducing the quality [131].

Another experiment that we did in this direction was the developing of a tool that is able to replace expensive equipment for eye-tracking studies [133, 134]. Furthermore, we organized a workshop at the MedivaEval 2015 and 2016 Benchmark Initiative [4] called "Context of Multimedia Experience", where we tried determine which movies are good for special watching situations like on a flight, based on content analysis [138]. The data that we used for these

[4] http://multimediaeval.org

experiments has been made publicly available [137]. During these studies, we also performed some quality of experience research to find out how good automatic quality metrics can mirror real users perception of quality [191, 42, 135, 157]. Furthermore, we collected a lot of data that can be useful for other researchers and released it as a publicly available dataset [132].

We also did experiments in the area of perception of deception[5], which focuses on the detection of manipulations of multimedia data like movies and images and on what intent these manipulations have been applied. Therefore, we collected a novel dataset [26]. We also helped to organize a workshop about this topic at the MediaEval 2015 and 2016 Benchmark Initiative called "Verification of Multimedia Use" [13]. The analysis of this data is ongoing work.

The main use case and the focus of the research is the field of medical multimedia. To be able to perform research in the medical field, two important requirements have to be fulfilled. Firstly, you have to be in contact with medical experts that are willing to share their domain knowledge with you, and secondly, you need to get data and annotations for this data. This is represented in more detail in the following sections.

## 2.4 Summary

It seems that medical multimedia systems are not in focus of research, and most of the research is focused on algorithms for the detection of diseases and not a complete system. The few examples that focus on more than one component seem to ignore the processing, or not reach the performance requirements. Most of the approaches in the GI tract use case are closer to medical image processing than multimedia research. The evaluation of the approaches is often based on a small amount of data, or the data is not publicly available to be comparable. Deep learning approaches seem to be a hot topic in the field of medical image processing, but because of their need for large amount of training data and that it is hard to understand how they make their decision, it can be problematic to use them in the medical scenario.

The search-based classification approach, which has been developed during this work and serves as the basis for the system, seems to be promising for our use case of disease classification in the GI tract. The main advantages of the search-based classification are that it is easy to understand, fast, accurate and easily expandable with other data types or use cases. To the best of our knowledge, the medical multimedia system researched in this thesis is the first that aims at total flexibility in terms of diseases that can be detected and at the same time focusing on the performance and a proper and comparable evaluation of it.

In the next chapters, we will present our medical multimedia system and all sub-components. Furthermore, we will show a complete evaluation of the system performance for accuracy and processing speed including the discussion of our GPU-based improvements of EIR.

[5]Perception of deception describes how manipulated content like images or videos is perceived from the users [25].

# Chapter 3

# The EIR System

An overview of the complete EIR system can be found in figure 3.1. Basically, EIR consists of three subsystems. The annotation subsystem collects and transfers knowledge and data from the medical experts into the system. The detection and automatic analysis subsystem consists of two parts. The detection part is responsible for detecting potential diseases in the current frame. The localization part uses the output of the detection part and tries to locate the disease in the image or frame. The visualization subsystem presents the output of the detection and analysis part to the medical expert for further analysis. The main purpose of EIR is to analyze multimedia data containing information about any parts of the GI tract in real-time. The goals for our system are:

- High disease detection accuracy.
- A complete pipeline for the whole system.
- Real-time processing for live support during colonoscopies.
- VCE and standard colonoscopy support by one system.
- Being expendable with different data types and diseases.

Achieving these goals came with a lot of challenges and required research in different directions like annotation, detection and visualization [142]. In the following, we will present a detailed description and overview of the complete EIR system and all the subsystems. This is followed by an experiment section that presents some of the experimental results that have not been published. Finally, we will present and discuss our efforts on the processing performance of EIR.

## 3.1 Annotation Subsystem

The main purpose of the annotation subsystem is to collect training data for the detection and automatic analysis subsystem. This type of data can only be collected with the help of medical experts. To make the collection process easy for the doctors and as efficient as possible, we combine manual annotations with automatic methods [145, 142].

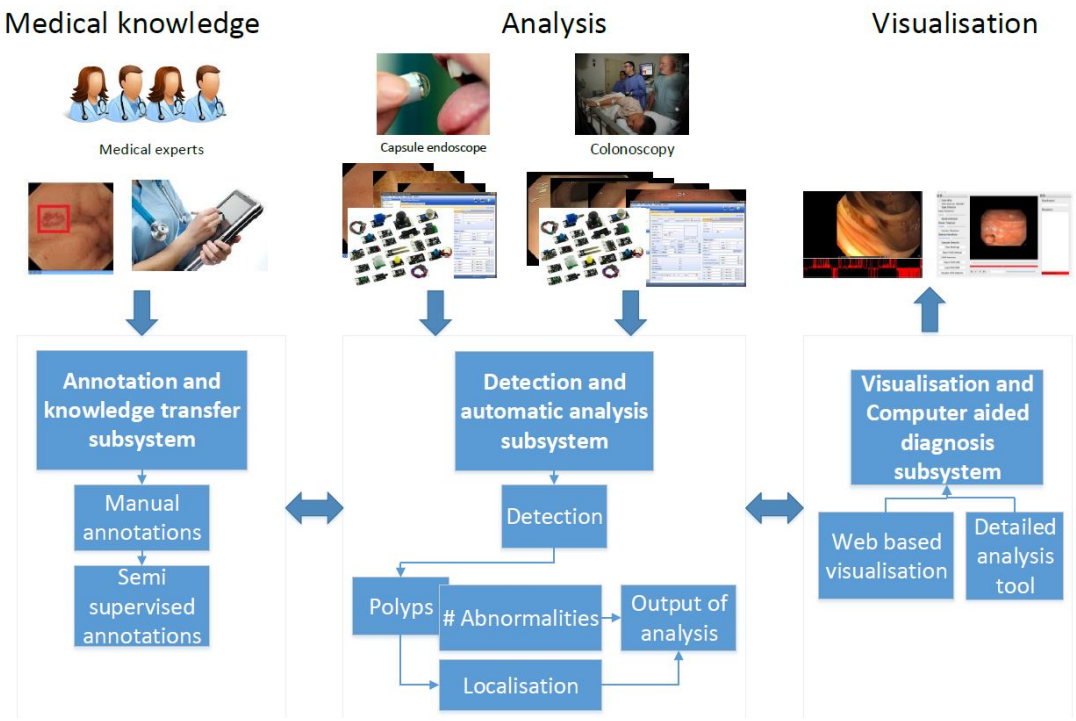



Figure 3.1: A complete overview of the EIR system. The system consists of annotation, detection and automatic analysis and visualization subsystems.

It is well known that training data is an important key factor to create a good classification system. Nevertheless, in the medical field, the time of experts and multimedia data are two resources that are quite limited. This is primarily the case because of high every-day workloads for doctors, but also due to legal issues. In many countries, patient consent has to be collected before research can be performed on patient related data, making data collection a very difficult 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 [44]. For example, in a VCE procedure, a video containing around $216,000$ - $1,000,000$ frames per examination is produced. An experienced endoscopist frequently needs one and even up to two hours to view and analyze all the video data [85]. Therefore, aside from 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 procedure documentation or teaching purposes. The current version of the annotation part consists of the semi-supervised annotation tool presented in [145, 4] and a cluster-based annotation tool [144].

### 3.1.1 Semi-supervised Annotation Tool

To reduce the amount of time doctors need to spend in the whole process, they only have to provide annotations in a single frame of the video or image. The specialist's knowledge is ideally only required for the first very basic identification of abnormalities and to tag them accordingly. The automatic step uses this information to track the regions of interest on previous and subsequent frames automatically. An example of the interface of the semi-supervised annotation tool can be found in figure 3.2.


Figure 3.2: The interface of the semi-supervised annotation tool. The user has only to mark the area with the disease and enter the name and a short description of it. The tool then automatically tracks the marked area and stores the examples in a training dataset.

The manual annotation part is performed by selecting regions of interest in a video sequence. The output from the manual annotation contains a single annotation for every region of interest in the video sequence or images. Using this information, object tracking algorithms are used in combination with manual corrections to generate a complete dataset. Most of the work in this step is done by the software. Depending on the quality of the video and the speed of camera movement, user intervention is needed to assure a high quality of tracking. A more detailed description of this part of the annotation subsystem can be found in [4].

There is of course still a fair amount of manual work involved to achieve good annotations. However, using a suitable tracking algorithm substantially reduce 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 [142]. 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. Every rectangle in such a list is described by the index of the video frame it belongs to, its position in pixel coordinates and its dimensions. The annotated frames are pooled together regarding their tags, which can be directly used in the detection and automatic analysis subsystem [145, 142].

### 3.1.2 Cluster-based Annotation Tool

To extended the annotation subsystem, we implemented a tool that allows the doctors to utilize global-features-based clustering to tag a large amount of data in a short period of time. The clusters are created based on visual global image features (described in detail in section 2.3.1) that are also used in our classification subsystem and the search-based algorithm. It is impor-

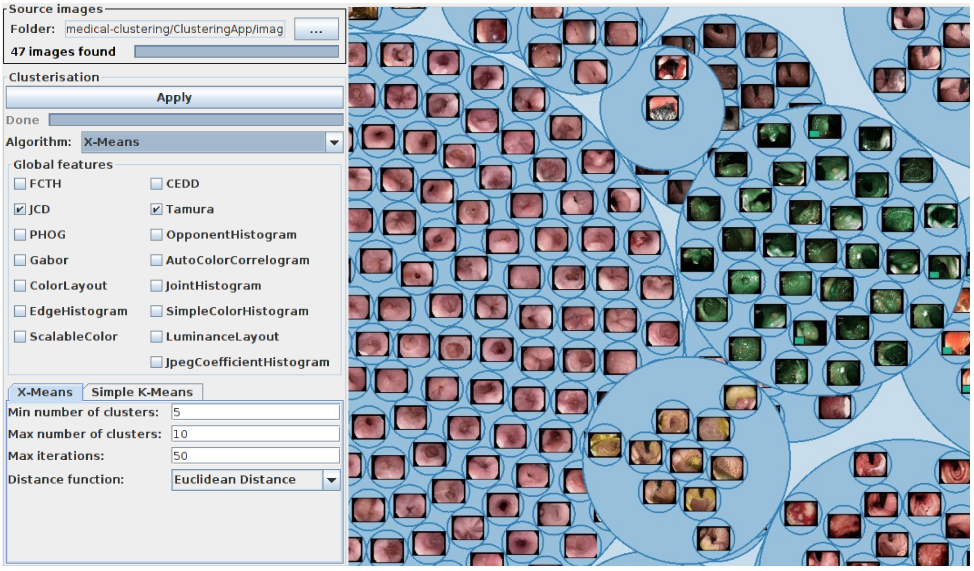



Figure 3.3: The interface of the clustering-based annotation application that makes the annotation process for medical personnel easier and more efficient [142].

tant to point out that the cluster-based annotation tool can be extended to support any type of multimedia data. After the clusters are created, the doctors can manually drag and drop images between different clusters and also annotate complete clusters at once. The clustering-based application has two main advantages. Firstly, it allows medical doctors to investigate and analyze vast collections data, for example, from endoscopic procedures by providing a configurable focus and context-view-based on frame similarity. Secondly, it makes it possible to utilize the focus and context-view for annotation of the dataset, making it more accessible for complimentary information systems such as our detection and automatic analysis subsystem [145, 142].

Figure 3.3 shows the annotation tool interface. On the upper left side, users can choose the folder containing the image collection. The clustering annotation tool supports at the moment unsupervised clustering (x-means clustering [124], the number of clusters does not have to be known in advance) and supervised clustering (k-means [69], the number of cluster has to be determined at the beginning), but can be easily extended with other clustering methods. The clustering algorithm can be selected in the settings in the lower left corner. For the clustering algorithm, the same features as supported by the detection subsystem can be chosen. If more than one feature is selected, they are combined using early fusion. In future versions of the tool, we will also support late fusion. The options at the bottom allow the user to specify the clustering parameters. These setting are set to default values recommended by literature [124, 69]. The clusters are represented using the clustered images. The closer an image is to the medoid of the cluster the closer it is to the center of the circle.

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 users can see information like the cluster center and the purity (which measures the ability of the clustering algorithm to recover known labels [72]) of the clusters by right clicking on the cluster. All images can be dragged and dropped between different cluster circles to improve the annotations. Finally, the medical experts can tag the

clusters, which adds the term used for the annotation to the name of the images in the cluster [145, 142]. The output of the clustering annotation tool is mainly used to identify and tag frames or images that contain abnormalities for the detection subsystem. Its output can also be used in the previous presented annotation tool to mark the exact position of abnormalities in the images which makes it also useful for the localization part.

## 3.2 Detection and Automatic Analysis Subsystem

The main purpose of the detection and automatic analysis subsystem is to automatically detect, analyze and localize endoscopic findings in the GI tract for standard colonoscopies and VCEs [145, 142].

The subsystem for detection and automatic analysis is designed in a modular way, so that it can be extended to different diseases or subcategories of diseases, as well as other tasks like size determination. At the moment, this subsystem consists of two parts, i.e., the *detection* part that detects irregularities in video frames and images, and the *localization* part that find the exact location of the disease in the image for frame. The detection part does not determine the location of an irregularity, but the localization part uses the output of the detection as input to perform localization on true positive frames only [142].

### 3.2.1 Detection

The detection part analyzes videos and images to find out if there is anything abnormal to be found. In our use case this means detecting diseases in the GI tract [145, 142]. All the frames that are processed in this part can be separated into two disjoint sets which can also be seen as the model for the classification algorithm. These two sets contain example images for abnormalities and images without any abnormalities. The detection is built in a modular way and can easily be extended with new models or submodels of different diseases. Such flexibility would, for example, allow to first detect a polyp and then distinguish between a polyp with low or high risk of developing into CRC by, for example, using the *NICE* classification[1]. Furthermore, it could be used to distinguish between different types of polyps to, for example, separate inflammatory, serrated and adenomatous polyps. To compare and determine the abnormalities in a given video frame (or image), global image features are used as described before. The main reason is because they are easy and fast to calculate, and because we are not interested in the exact position at this point of the system. A detailed overview of the steps within the detection that are described in the following text, can be found in figure 3.4.

For the implementation of the search-based classification of the detection part we used the Lire [96] open source library for content-based image retrieval as a starting point. 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.

Lire uses Lucene[2] indexes for storing and searching image feature data [171]. Lucene indexes are structured in documents, fields and terms. An index contains a sequence of docu-

[1] `http://www.wipo.int/classifications/nice/en/`
[2] `https://lucene.apache.org/`

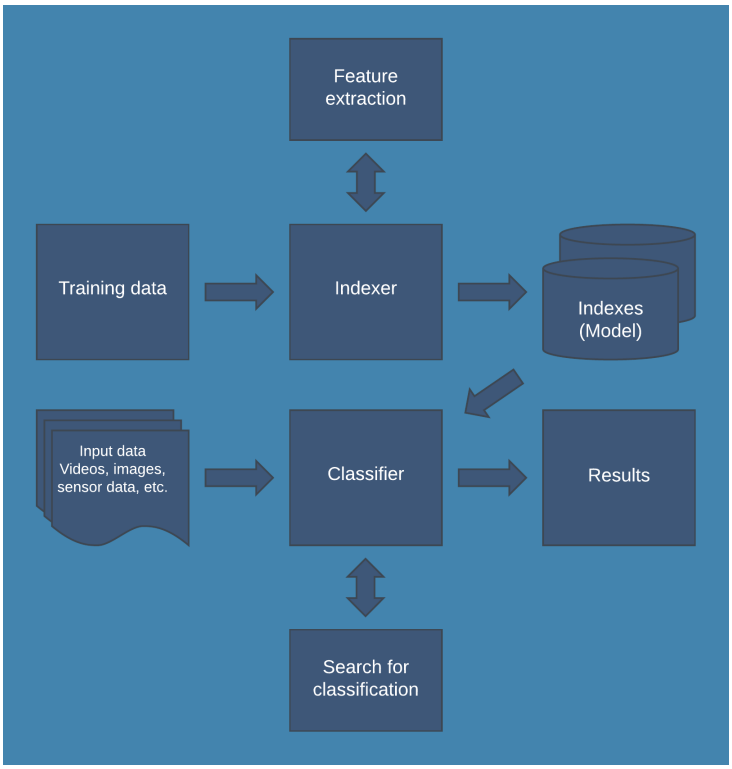



Figure 3.4: This diagram shows the detailed steps performed by the detection part of EIR. The training data is first indexed by the indexer. The indexer indexes different types of features from the input data, which are extracted in the feature extraction part. The indexes are used by the classifier as a model to classify input data. The classifier performs a search-based classification on the data to get the final results, which then can be used for the localization determination or presented to the user.

ments, where a document is a sequence of fields, a field is a sequence of terms and a term is a string [171]. Indexes that are used as the model of the search-based classification are created using as many example frames as possible, but it is important to point out that the number of needed examples is rather low compared to other methods like deep learning. 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 results, the detection subsystem then decides which abnormality (if any) the input frame belongs to [142].

The complete classifier is realized using two separate tools, an indexer and a classifier. The indexer and the classifier can be found in our open source project *OpenSea* [104]. The main task 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 then used as input data for the search-based classifier. The indexer is created as a separate tool, and in a way that it is easy to distribute it on heterogeneous architectures. 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 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 disease require different set 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 subdirectory inside the indexed directory. If multiple directories are passed for indexing, it creates a separate index for each directory [145, 142].

The classifier can be used to classify video frames from an input video into as many classes as the detection part 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. 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 [142].

The classifier also creates an HTML page with a visual representation (see figure 3.5) of the results. Once the processing is finished, a benchmarking function outputs evaluation information (bottom part of figure 3.6(a)).

To be able to use the benchmark function, the input data indexes must contain either negative or positive samples only, or must have the sample type encoded in the file names of the indexed images. The classifier is parallelized, and it can be chosen how many cores should be used to process the data. A GPU implementation has also been developed, and experiments show that performance improved further (presented in detail in section 3.5) [145, 142].

### 3.2.2 Localization

As mentioned before, the detection part does not determine the location of the detected irregularities. The location determination is performed by the localization part of the detection and automatic analysis subsystem, which uses the output of the detection as input. An overview of the detailed steps performed by the localization can be found in figure 3.7.

Currently, the localization supports finding the location of polyps, but the localization is also built in a way that it can be extended to any other automatic detectable diseases in our system. The exact positions could be used for live examinations, archiving, and future size determination. The processing of the images is implemented as a sequence of intra-frame pre- and main-filters (see figure 3.6(b) for a console output example). Pre-filtering is needed because we use local image features to find the exact position of objects in the frames, which is easier on pre-filtered images.

Disease objects or areas itself can have different shapes, textures, colors and orientations. They can be located anywhere in the frame and can also be partially hidden or covered by biological substances like seeds or stool. Furthermore, endoscopic findings can be lighted by direct or reflected light. Moreover, the image itself can be interleaved, noisy, blurry and over-/under-exposed, and it can contain borders and sub-images (from the colon location function that some colonoscopy processors provide to show where in the colon the endoscope is). Additionally, it can have various resolutions depending on the type of endoscopy equipment or VCE used.

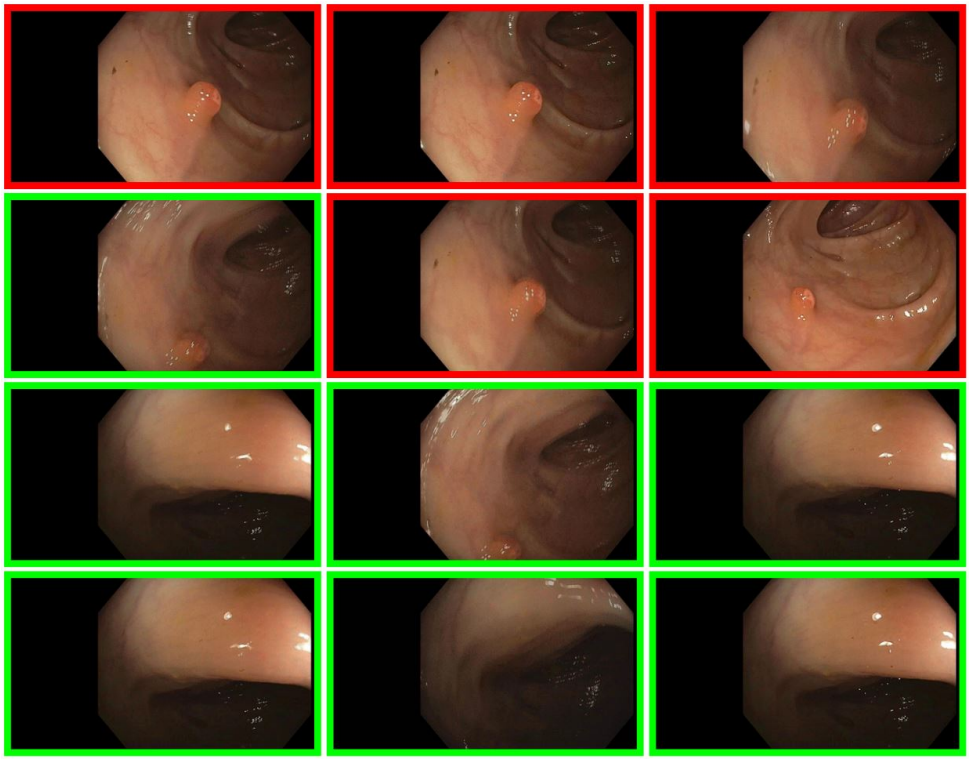


Figure 3.5: Result output of the detection part using the features JCD and Tamura. One can see that the detection part could almost always find the polyp containing frames. The first image on the second row is an example for a false negative result [142].

Endoscopic images usually have a lot of flares and flashes caused by the high power light source on the endoscope. All these nuances can negatively affect the local features detection methods and have to be treated specially to reduce the impact on the localization precision.

In our case, we used several sequentially applied filters to prepare raw input images for the following analysis. These filters are RGB to YCbCr color space conversion, removal of borders and sub-images, flares masking and low-pass filtering. After the pre-filtering, a local-feature-based analysis follows that uses the pre-processed images as input [145, 142].

As described above, we have implemented the localization of colon polyps using a local features approach. The main idea of this 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 underlying surface with legs varying in their thickness. The locations of these polyps can be approximated using elliptical shape regions that consists of local features that differ from the surrounding tissue with a high probability. In the localization algorithm, we use the following sequence of filters: binary noise reduction filter, 2D-gradient filter, threshold borders detection filter and binary noise removing filter. After this filtering, the next step creates a filtered binary image, approximated by a set of ellipses from which we built energy maps based on the ellipsis size and border points precision approximation and matching.

The final coordinates of one or more polyps in the frame are chosen by searching for the maximums in the energy map. At the moment the algorithm gives as an output four locations. These four possible locations are the ones with the highest probability to be on a polyp. To

```
using 4 threads for classifying.
...
image_hortVD_wp_68_71.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_8.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_79.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_82.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_78.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_77.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_81.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:NEGATIVE
image_hortVD_wp_68_83.jpg -> Tamura:NEGATIVE LateFusion:NEGATIVE JCD:POSITIVE
image_hortVD_wp_68_80.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_9.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_102.jpg -> Tamura:POSITIVE LateFusion:NEGATIVE JCD:NEGATIVE
image_hortVD_wp_68_84.jpg -> Tamura:NEGATIVE LateFusion:NEGATIVE JCD:POSITIVE
image_hortVD_wp_68_101.jpg -> Tamura:POSITIVE LateFusion:NEGATIVE JCD:NEGATIVE
image_hortVD_wp_68_103.jpg -> Tamura:POSITIVE LateFusion:NEGATIVE JCD:NEGATIVE
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Feature     TP   TN  FP  FN  Precision Recall    TNRate    FPRate    Accuracy  FMeaseure WFMeasureMccMeasure
Tamura      165  34  37  21  0.816832  0.887097  0.478873  0.521127  0.774319  0.850515  0.850515  0.399001
LateFusion  172  51  20  14  0.895833  0.924731  0.718310  0.281690  0.867704  0.910053  0.910053  0.661481
JCD         162  45  26  24  0.861702  0.870968  0.633803  0.366197  0.805447  0.866310  0.866310  0.509303

writing html output to: results-1435447137.html
duration: 78.293seconds.
```

(a) Console output of the detection part using the features JCD and Tamura.

```
user@user-media-dep: ~/EIR

user@user-media-dep:~/EIR$ ./localizer.sh
CPUs: 4
Using 4 threads
Display 0: 2048x1152
Processing images...
  src_id ;obj_n_lim;___TP___;___TN___;___FP___;___FN___;precision;  recall ; F-score ;rejected
      all;         1;     1433;        0;     3035;     3069;   0.3207;   0.3183;   0.3195;    15046
      all;         2;     1981;        0;     6791;     2521;   0.2258;   0.4400;   0.2985;    15046
      all;         3;     2333;        0;    10754;     2169;   0.1783;   0.5182;   0.2653;    15046
      all;         4;     2617;        0;    14798;     1885;   0.1503;   0.5813;   0.2388;    15046
      all;         5;     2784;        0;    18975;     1718;   0.1279;   0.6184;   0.2120;    15046
      all;         6;     2900;        0;    23215;     1602;   0.1110;   0.6442;   0.1894;    15046
      all;         7;     2995;        0;    27471;     1507;   0.0983;   0.6653;   0.1713;    15046
      all;         8;     3075;        0;    31774;     1427;   0.0882;   0.6830;   0.1563;    15046
---
 src_size ;frame_time ms; n_frames ; time_total s
       all;      945.302;     19514;    18446.627
   384x288;      824.840;       612;      504.802
   712x480;        0.000;     13500;        0.000
   856x480;     2633.184;      2435;     6411.804
   960x540;     3886.087;      2967;    11530.021
Done
Processing time 4754.769 sec
user@user-media-dep:~/EIR$
```

(b) The final output of the localization part for polyps. The localizer can easily be extended to localize different disease detect by the detection part.

Figure 3.6: System output for the detection and localization part after the analysis. It includes general results per frame and all evaluation metrics that are provided by the system [142].

further improve the results, a method that will make it possible to determine which one of the four positions has the highest probability to be on a polyp will be needed. An example for the output of the localizer is shown in figure 3.8. Correct locations are marked with green crosses and incorrect locations with red crosses. The area of the polyp in the image is marked with transparent blue. The second image in the first row shows an optimal result, the last image in the second row shows a not optimal output with only one cross on the polyp [145, 142].

## 3.3 Visualization Subsystem

The purpose of the visualization subsystem is to provide the results of the automatic detection and analysis subsystem to the medical experts who are supposed to use the output of it for computer aided diagnosis. It is important to point out that the visualization part was not a main

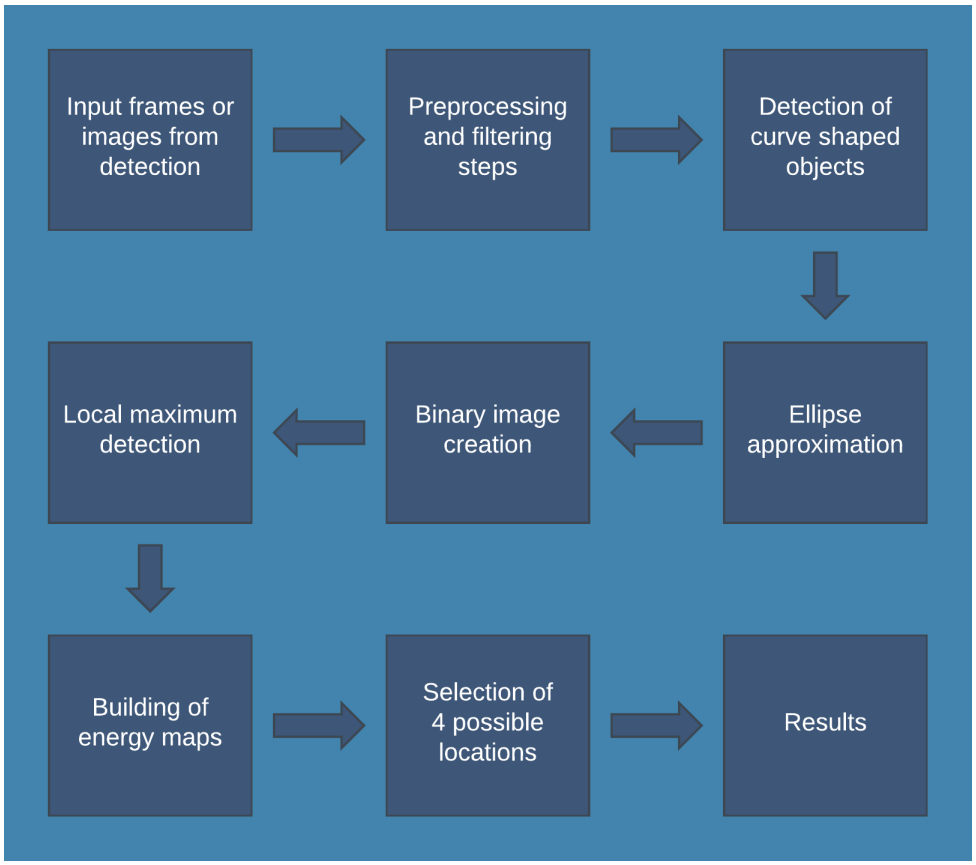



Figure 3.7: This diagram gives an overview of the most important steps performed by the localization part of EIR. The localization receives input frames or images containing polyps from the detection part. The input frames are preprocessed and filtered. In the filtered images curve shaped object detection is performed. This is followed by ellipse approximation and binary image creation. The output of this is then used to perform local maximum detection and building of energy maps. The energy maps are then used to select four locations that are most probably showing a polyp.

focus of this thesis and mainly implemented to allow medical experts to use the output of the EIR system. 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 experts use a significant part of their time on this task, and they are therefore in need of multimedia systems that can support them by minimizing errors and increase the efficiency in this process.

For EIR, we developed two different tools that can be used to visualize the output of the automatic analysis to the medical experts. The first one is based on the semi-supervised annotation tool [4], which was described in section 3.1.1. We extended the tool with the possibility to load and visualize the output of the detection and automatic analysis subsystem. An example for how the visualization looks with this tool can be seen in figure 3.9. This version is very convenient for research purposes because of the additional image processing functions (e.g., specular highlight filter or border detection), but it can be too complicated to use for medical personnel in the hospital.

The second tool for the visualization that we developed is a web-based application. The goal

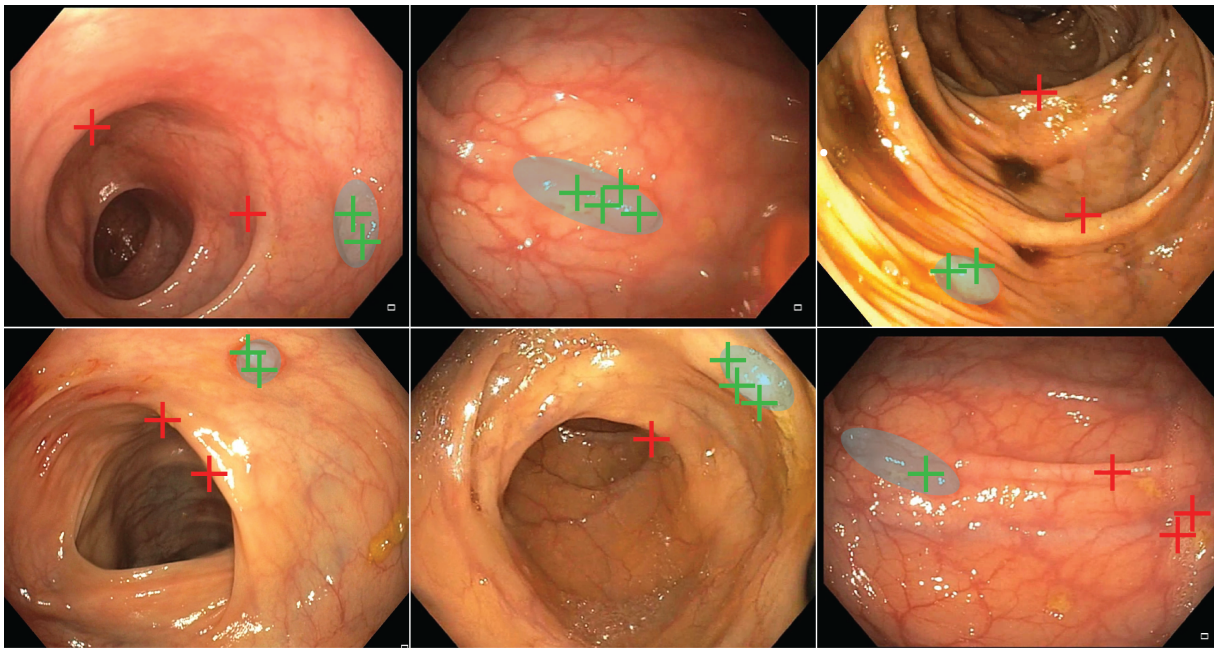


Figure 3.8: Output of the localization part marking the four possible locations of polyps determined by the algorithm. True positives are marked with green crosses, false positives are marked with red crosses. The exact area of the polyp is highlighted with transparent blue. The algorithm gives four possible locations for a polyp in the frame. For future work, this will be reduced to one cross per frame (the one with the highest probability to mark a polyp) [142].

of the web-based visualization tool is to support medical experts with an application that is easy to use and makes it easy to share data amongst participating medical experts in the future. The interface and an example visualization are shown in figure 3.10. Our prototype facilitates the output of the detection and localization part and creates a web-based visualization which will be combined in the future with a video sharing platform [55] 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 web-based 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 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 [144, 145, 139, 142].

The prototypes for the visualization subsystem can 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 use the output of our system and share data with other experts. Finally, not just the visualization applications are important but also an understanding of how humans perceive multimedia content and how different aspects of the content influence them differently [139], which is out of scope of this work and an important task for future work.

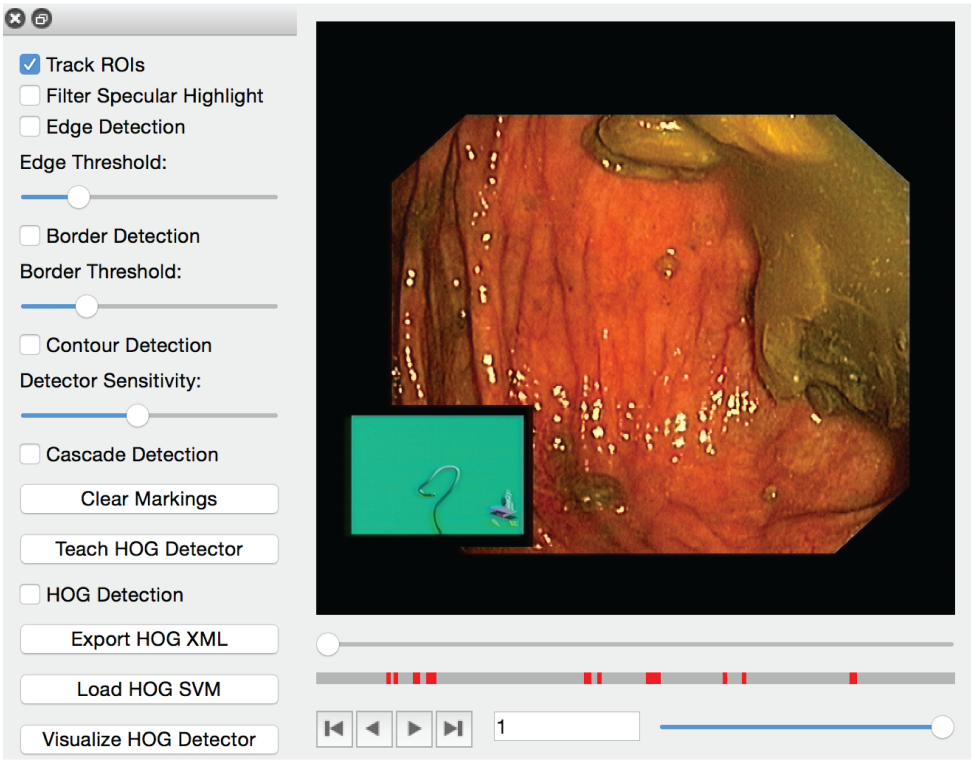



Figure 3.9: Visualization of the output of the automatic analysis subsystem of EIR using an extended version of the semi-supervised annotation tool. The time line below the videos indicates with red color where significant findings are located in the video. The tool also provides additional image processing functionalities such as filtering of specular light and edge detection.

## 3.4 System Evaluation

In this section, we will present the initial experiments that we conducted on the EIR system for our polyp detection use case. We tested the whole system in terms of detection accuracy and system processing performance. The requirements of the system that we are evaluating are, (i) reaching real-time performance (being able to process 25-30 frames per second) and (ii) achieving high detection accuracy (at least equal to the best related approaches in table 2.1, above 90% for precision, recall/sensitivity and F1 score). We also participated in the Endovis Automatic Polyp Detection in Colonoscopy Grand Challenge at the 2015 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) with our system. Furthermore, we present results of the EIR system for CPU-based and GPU-based distribution of the workloads (CPU-based in section 3.4.4 and GPU-based in section 3.5). This is followed by a proof-of-concept for the multiple disease detection functionalities of EIR in section 3.6.

For any of the subsequent measurements, we used the same computer which was an *old* 32 AMD CPU cores Linux server with 128 GB ram. This evaluation is similar to the one conducted for our publication [139] and [142]. Nevertheless, this version differs in the used hardware and the presentation.

As mentioned previously, we used the ASU Mayo Clinic polyp dataset [168] as training and test data. This dataset is the biggest publicly available dataset consisting of 20 videos,

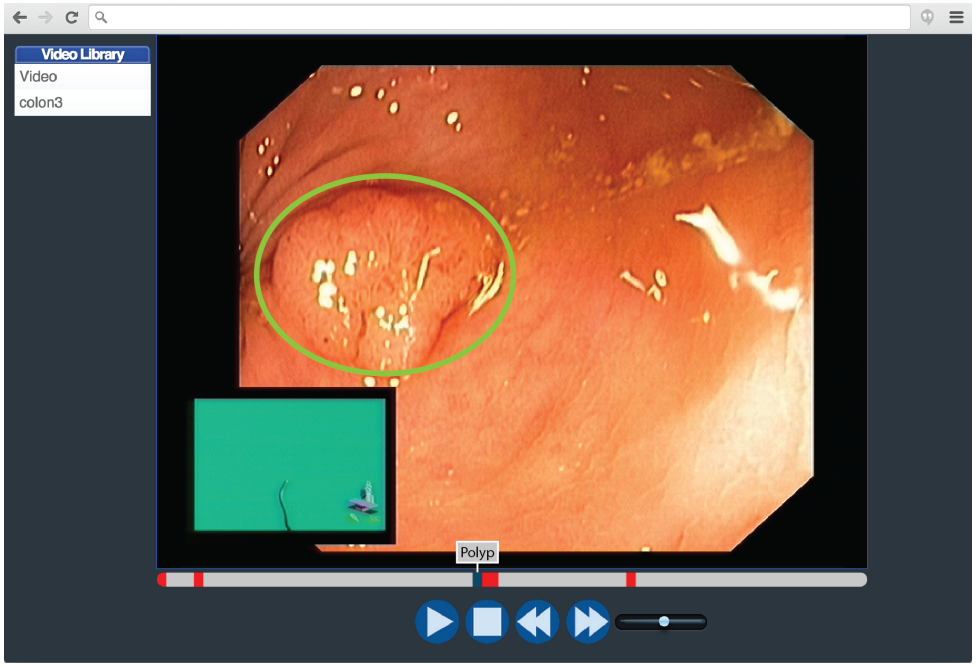



Figure 3.10: A web-based visualization tool for the output of the automatic analysis subsystem of EIR. The application is built in a way that it is easy to use and expandable with sharing features in the future. The red color in the time line shows where significant endoscopic findings are located with a tag on top of it naming the finding. The location is marked with a circle around the disease.

converted from WMV to MPEG-4 for the experiments, with a total number of $18,781$ frames and resolutions between $720 \times 480$ to $1,920 \times 1,080$ pixels. The videos are named using *wp_* or *np_* plus the number of the video. A complete overview of all videos and their length, resolution and if they contain a polyp (if a video contains a polyp, it only contains a single one) or not can be found in table 3.1.

### 3.4.1 Detection Accuracy

For all detection and localization accuracy experiments, we used the common standard metrics precision, recall/sensitivity and F1 score. Precision is computed by true positives ($tp$, correct as positive classified) divided by $tp$ plus false positives ($fp$, false classified as positive class). The higher the precision is, the more precise is the detection algorithm.

$$Precision = \frac{tp}{tp + fp}$$

Recall (also called sensitivity) is computed by $tp$ divided by $tp$ plus false negatives ($fn$, false classified as negative class). A high recall shows that the algorithm is able to detect all occurrences of polyps.

$$Recall = \frac{tp}{tp + fn}$$

In an optimal case both, $precision$ and $recall$, are high. Usually, precision and recall are in such a relation that a higher precision leads to a lower recall and vice versa [129]. A way to cal-

| Video | Length in minutes | Resolution | Contains polyps |
|---|---|---|---|
| np_5 | 00:22 | $720 \times 480$ | no |
| np_6 | 00:27 | $720 \times 480$ | no |
| np_7 | 00:25 | $720 \times 480$ | no |
| np_8 | 00:23 | $720 \times 480$ | no |
| np_9 | 01:01 | $720 \times 480$ | no |
| np_10 | 01:04 | $720 \times 480$ | no |
| np_11 | 00:51 | $720 \times 480$ | no |
| np_12 | 00:58 | $720 \times 480$ | no |
| np_13 | 01:00 | $720 \times 480$ | no |
| np_14 | 00:54 | $720 \times 480$ | no |
| wp_2 | 00:10 | $1920 \times 1080$ | yes |
| wp_4 | 00:30 | $1920 \times 1080$ | yes |
| wp_24 | 00:17 | $720 \times 480$ | yes |
| wp_49 | 00:16 | $856 \times 480$ | yes |
| wp_52 | 00:36 | $856 \times 480$ | yes |
| wp_61 | 00:11 | $1920 \times 1080$ | yes |
| wp_66 | 00:13 | $856 \times 480$ | yes |
| wp_68 | 00:08 | $1920 \times 1080$ | yes |
| wp_69 | 00:20 | $1920 \times 1080$ | yes |
| wp_70 | 00:13 | $856 \times 480$ | yes |

Table 3.1: Overview of all videos used for the experiments. For each video name, resolution and polyp occurrence is reported.

culate the quality of a classification system that considers both measures is the F1 score, which is the harmonic mean between precision and recall and gives an idea of the overall performance of the system.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Moreover, we decided to use leave-one-out cross-validation to evaluate this part of the system. Leave-one-out cross-validation (cross testing of different splits of the data, reported is the average performance over all splits) is well suited to show generalization potential and robustness of a predictive model [130]. Therefore, 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. This process is repeated until every item or portion has been used for testing exactly once [33]. The EIR system allows us to use several different global image features for the classification. All features can also be combined using late fusion. The more image features we use at the same time, the more computationally expensive the classification becomes. Furthermore, not all image features are equally important or provide equally good information for our use case. As a first step, we therefore need to determine which image features are suited best for the detection.

To be able to understand which image features provide the most information gain, we generated indexes containing all possible image features for all frames of all video sequences for

our dataset. We then used these indexes for several different measurements and also for leave-one-out cross-validation (leaving one video out from the training data at the time and using it as test data, repeating the procedure until each video had been left out once).

The built-in evaluation functionality of EIR provided information on the performance of different image features for benchmarking. The output of the evaluation function provides separate information for every single image feature, as well as the performance for late fusion of all the selected global image features. As mentioned in section 2.3.5, we did not use PCA for this step because we use late fusion.

The obtained values for true positives, true negatives, false positives and false negatives for all the runs are used to calculate the metrics for the leave-one-out-cross validation.

The results of this first test are presented in table 3.2 [142]. The global image feature that generally achieves the best results is CEDD [96]. All features used here have been described in detail in section 2.3.1. The table also reveals that the image features JCD, EdgeHistogram, Rotation Invariant Local Binary Patterns, Tamura and Joint Histogram achieve similar positive results. Late fusion of all used image features achieves the best results.

Nevertheless, it is not feasible yet to use late fusion of all image features for the classification because the calculation, indexing and searching of all image features is computationally expensive. After some more testing and also based on our findings in [136], we decided that a late fusion of two image features will provide optimal results in terms of detection accuracy and at the same time minimizing the computational requirements for our use case. Furthermore, some of the two feature combinations reached accuracy at least equal or better compared to the state-of-art presented in this work. Combinations of more than two features is therefore not necessary for our use case because it will not increase the performance, but it can be an interesting task for future versions of EIR.

To determine which features work best for this fusion, we ran another experiment where we tested all, by the EIR system supported, possible combinations of features. For these tests, we performed on only one video to avoid overfitting on the dataset (the classifier adjusts to very specific random features in the dataset, which happens especially if training data is rare, or learning is performed too extensive on the data [59]).

One can see that many combinations perform well and can be used, and the detection accuracy shows only minor differences. Nevertheless, based on the results for this evaluation presented in table 3.3 and the general characteristics of the used global features, we decided to use the features JCD and Tamura for our use case. The reason for this decision is that they have a good trade-off between detection accuracy and processing performance and at the same time, both features combined, result in a $186$ dimensions feature vector, which makes them easy and fast to compute [142].

After we found the best combination of two features, we evaluated the classification performance using these two image features. Using the two features, we conducted again a leave-one-out cross-validation with all available video sequences, and the results are presented in table 3.4. Using JCD and Tamura for late fusion, we achieved an average precision of $0.889$, an average recall of $0.964$ and an average F1 score value of $0.916$. Since the used dataset is not balanced in terms of how many negative and positive examples we have and also how many frames each video contains, we calculated also weighted metrics. If we weight the values contributed by every single video with the number of frames in the video, we achieve an average precision of

| Feature | True Positive | True Negative | False Positive | False Negative | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|---|
| JointHistogram | 3,369 | 13,826 | 1,085 | 511 | 0.7563 | 0.8682 | 0.8084 |
| JpegCoeffHist. | 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 |
| FuzzyOppHist. | 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 |
| RotInvtLlBP | 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 |
| LocBinPattAOpp | 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 |
| RankAndOpp | 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 |
| EdgeHistogram | 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 |

Table 3.2: Leave-one-out cross-validation for all, by the EIR system supported, features [142].

$0.9388$, an average recall of $0.9850$, and an average F1 score value of $0.9613$. These results proof that it is possible to detect polyps with a precision of around $94\%$ and almost $99\%$ of all polyp containing frames are detected by the EIR system, which can be seen as very good [142].

### 3.4.2 Localization Accuracy

For the experiment to test the localization accuracy, we used the ground truth provided in the ASU Mayo clinic polyp dataset. Furthermore, we also used the CVC-ClinicDB dataset [10] containing $612$ still images from $29$ different videos from colonoscopy examinations that were used during the MICCAI challenge. Both datasets' ground truth contains the exact positions of the polyps in the frame as binary images where the polyp location is presented by zeros. The CVC-ClinicDB dataset contains only images with polyps and none without. Table 3.5 shows the performance of the localization part of EIR. The first row shows the results for the CVC-ClinicDB dataset, all following rows show the performance for videos of the ASU Mayo dataset.

The performance is presented for every video and also as average. We did not include videos that did not contain a polyp at all because they are filtered out by the detection part beforehand, and we counted the output of the localization as positive if the algorithm marked a location within the area of a polyp on the binary image. The localization part has a precision of $0.3207$, a recall of $0.3183$ and a F1 score of $0.3195$. To determine the recall in terms of finding the exact location of the polyp, the false positives have also to be counted as false negatives (because the

| Feature Combination | True Positive | True Negative | False Positive | False Negative | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|---|
| Rot.Inv.LBP & Tam | 162 | 22 | 153 | 0 | 0.5142 | 1 | 0.6792 |
| PHOG & Tam | 161 | 23 | 152 | 1 | 0.5143 | 0.9938 | 0.6778 |
| JpegCoeff.Hist. & Tam | 162 | 21 | 154 | 0 | 0.5126 | 1 | 0.6778 |
| Gabor & Tam | 162 | 20 | 155 | 0 | 0.5110 | 1 | 0.6764 |
| FuzColHist. & Tam | 162 | 18 | 157 | 0 | 0.5078 | 1 | 0.6735 |
| FuzOppHist. & FuzColHist. | 160 | 17 | 158 | 2 | 0.5031 | 0.9876 | 0.6666 |
| JCD & OppHist. | 135 | 67 | 108 | 27 | 0.5555 | 0.8333 | 0.6666 |
| JoiHist. & JpegCoHist. | 162 | 12 | 163 | 0 | 0.4984 | 1 | 0.6652 |
| ColLay & FuzColHist. | 162 | 11 | 164 | 0 | 0.4969 | 1 | 0.6639 |
| FuzColHist. & JoiHist. | 162 | 11 | 164 | 0 | 0.4969 | 1 | 0.6639 |
| FuzOppHist. & JoiHist. | 162 | 11 | 164 | 0 | 0.4969 | 1 | 0.6639 |
| FuzOppHist. & SimColHist. | 162 | 11 | 164 | 0 | 0.4969 | 1 | 0.6639 |
| JoiHist. & RotInLBP | 162 | 11 | 164 | 0 | 0.4969 | 1 | 0.6639 |
| JoiHist. & SimColHist. | 162 | 11 | 164 | 0 | 0.4969 | 1 | 0.6639 |
| FuzOppHist. & Gabor | 161 | 13 | 162 | 1 | 0.4984 | 0.9938 | 0.6639 |
| JCD & JpegCoHist. | 161 | 13 | 162 | 1 | 0.4984 | 0.9938 | 0.6639 |
| CEDD & FuzColHist. | 159 | 17 | 158 | 3 | 0.5015 | 0.9814 | 0.6638 |
| JpegCoHist. & RotInLBP | 152 | 31 | 144 | 10 | 0.5135 | 0.9382 | 0.6637 |
| **JCD & Tam** | 162 | 10 | 165 | 0 | 0.4954 | 1 | 0.6625 |
| CEDD & Tam | 162 | 10 | 165 | 0 | 0.4954 | 1 | 0.6625 |

Table 3.3: Top 20 results of the performed experiments for late fusion. Each combination contains two image features for the video wp_61, sorted by F1 score [142].

localization algorithm in the current state cannot not determine if their is a polyp in the frame or not). These results indicate that the localization works as intended but has large potentials for improvement. One problem that can be interesting to solve for future improvements is that the localization actually outputs four possible disease positions per frame, which should optimally be reduced to one location with the highest change to be on the polyp. In almost all cases, one of the four possible locations points at a polyp, but for the evaluation all four points where included in the calculations, which influences the performance metrics negatively (which is depicted in the large number of false positives).

### 3.4.3 MICCAI Challenge Results

To see how our method compares to other state-of-the-art methods, we participated in the MICCAI challenge (Endovis Automatic Polyp Detection in Colonoscopy Grand Challenge at the 2015 International Conference on Medical Image Computing and Computer Assisted Intervention). The challenge consisted of two parts. The first part (i), was the polyp localization, where the task was to find out if the proposed method could cope with important polyp appearance variability and, therefore, accurately determine the location of the polyp in a frame. The second part (ii), asked the questions if the proposed 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 it could be detected.

| 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 |
| Average: | | | | | 0.8890 | 0.9640 | 0.9160 |
| Weighted average: | | | | | 0.9388 | 0.9850 | 0.9613 |

Table 3.4: Performance evaluation by leave-one-out cross-validation for all available videos, using JCD and Tamura features combined via late fusion [142].

We did not expect very good results for the challenge since EIR is not built only for polyp detection. The other participants used a wide range of different methods to detect polyps and were more specialized in the topic. These methods ranged from hand crafted features like contour- or shape-based detection used in combination with traditional machine learning approaches to neural networks. We identified several interesting challenges that come with polyp detection during the challenge such as blurry images due to camera motion, size differences, lighting and objects that look like polyps but are not, such as contaminants [142].

Table 3.6 shows the result for the polyp localization part based on the CVC-ClinicDB dataset. EIR was on the fourth place out of six. Details about the implementation of the other participants' methods are not available, but the RUS approach used a deep learning method. Based on the fact that our system is not built for only polyp detection, the results are very 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 3.7 gives and overview of the results for the detection latency part. For the latency, EIR performed second best out of all participants. This is a very good result, and a positive confirmation about the real-time performance compatibility 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.

| Data set | 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 |

Table 3.5: Performance evaluation of the localization algorithm [142]. To be able to determine the true recall in terms of finding the exact location of the polyp, the false positives have also to be counted as false negatives (because the localization algorithm in the current state cannot not determine if their is a polyp in the frame or not).

| 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 3.6: Results of the MICCAI polyp localization challenge [142].

| 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 3.7: Results of the MICCAI polyp detection challenge. The table shows the detection latency in milliseconds and F1 score [142].

Overall, the results of the challenge are positive for a system that is designed to be expandable with different diseases and use cases. We proved that we can compete and outperform

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

Table 3.8: Performance evaluation of the indexing part. Four different datasets with different sizes have been tested to show the scaling capability of the indexing part [142].

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 [142]. It is also important to point out that we participated in the MICCAI challenge with an early version of EIR and that the results might be even better with the current version.

### 3.4.4 System Processing Performance

As discussed before, an important requirement for medical multimedia system is scalability and processing performance. Since the use case is to use the system to do mass screening for lesions in the GI tract, using video sequences recorded live with colonoscopy or VCEs, we need both scalability and fast processing. For the processing performance evaluation, we decided to use the same configuration of the EIR system as in the accuracy performance evaluation, which is the late fusion of JCD and Tamura. It is important to reach real-time performance in terms of processing a video and reach a frame rate of not less than $25$-$30$ FPS. For all tests, we used three videos from three different endoscopic devices and different resolutions. The three videos are wp_4 with a resolution of $1,920 \times 1,080$ and $910$ frames, wp_52 with $856 \times 480$ and $1,106$ frames and np_9 with $712 \times 480$ and $1,843$ frames (see table 3.1). We chose these videos to show the performance under different requirements that the system will have to face when it is used in a real hospital [142].

#### 3.4.4.1 CPU Processing

To test how the different parts of our system scale in terms of used CPU cores, we performed several tests on our test machine. For all tests, we measured time per frame for the number of used cores. We also conducted some experiments to understand the influence of the size of the training data on the performance.

**Processing Performance for Indexing:** For the detection approach, we first measured the indexing part that creates the model that is later on used by the classifier. This process does not have to be in real-time and can be seen as batch processing, but it should at least be scalable for larger datasets. We did all experiments on the AMD Linux machine described earlier using $16$ CPU cores. Extracting two features and indexing them for the whole ASU Mayo dataset takes in average $8$ milliseconds per frame. There is no big difference between the indexing time for different resolutions. We tested the scaling potential by indexing different datasets. The first dataset ($D1$) contains $3,871$ frames, the second ($D2$) contains $14,909$ frames, the third ($D3$) contains $29,818$ frames and the last ($D4$) contains $100,000$ frames. Table 3.8 gives an overview

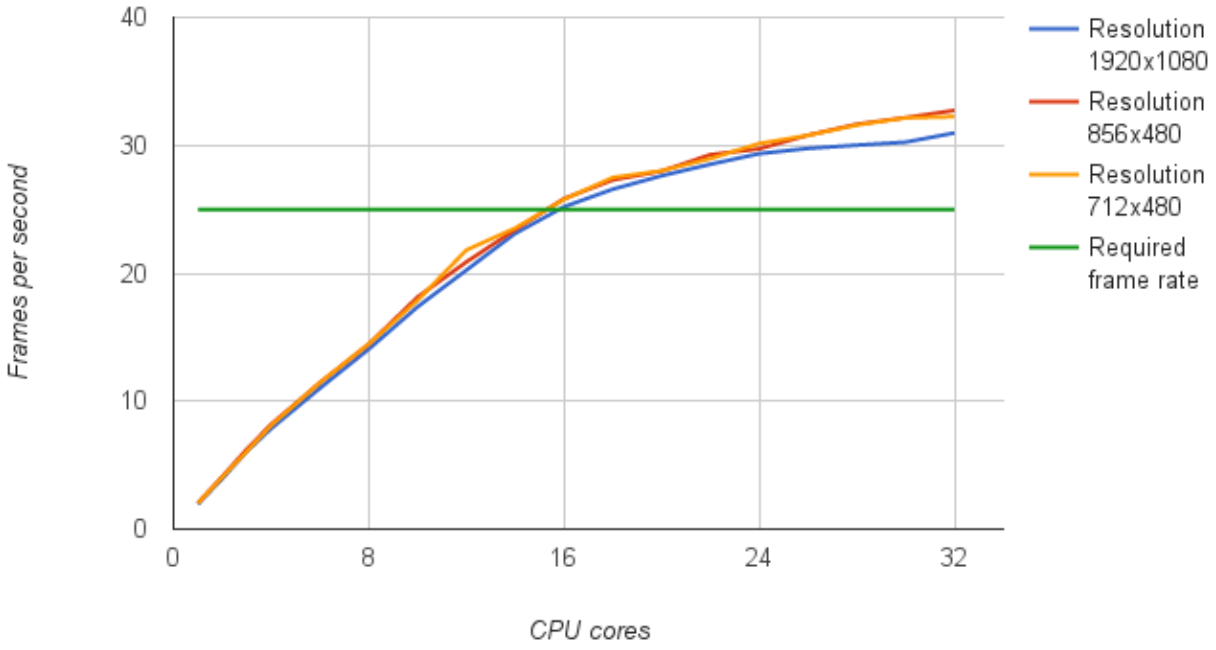



Figure 3.11: Detection performance in terms of FPS depending on the number of CPU cores and the resolution of the videos. The videos are wp_4 with a resolution of $1,920 \times 1,080$, wp_52 with a resolution of $856 \times 480$ and np_9 with a resolution of $712 \times 480$. For all videos, we observe that the required frame rate is reached with $16$ CPU cores used in parallel.

of the overall results. Our most interesting finding is that a larger dataset leads to a faster indexing time per frame. We assume that this happens because of the Java runtime optimizer. Furthermore, we could not measure an increase of the processing time after more than $30,000$ frames in the dataset. We think that the limiting factor is the I/O part since increasing the number of CPU cores did not increase performance. The experiments on the indexer part reveal that the indexer is able to deal with larger datasets, and it should be able to meet all requirements of the system for future tasks.

**FPS Performance:** Next, we tested the performance of the detection and localization parts in terms of processing speed. This is an important factor, since the system should provide a result as fast as possible and not slower than $25$ FPS making it usable for live applications. For all tests, we used the three different videos described before.

*FPS Performance for Detection:* Figure 3.11 shows the detection part performance for the three tested resolutions in FPS. The required FPS for all three videos are reached with $16$ CPU cores used in parallel.

*FPS Performance Localization:* Figure 3.12 shows the performance of the localization part for FPS. For the highest resolution, namely $1,920 \times 1,080$, the best result is $7.9$ FPS. A significant code optimization and using of GPU for accelerated calculations will be needed to reach required FPS. For resolution $856 \times 480$, the required FPS are almost reached with $32$ CPU cores used in parallel. The best result is $22.5$ fps for this resolution. A code optimization will be needed to reach the required FPS. For the final video with the resolution $712 \times 480$, the required FPS are reached with $19$ CPU cores in parallel. The outcome of these experiments for the localization part clearly shows that our system can reach real-time requirements but needs some optimization.

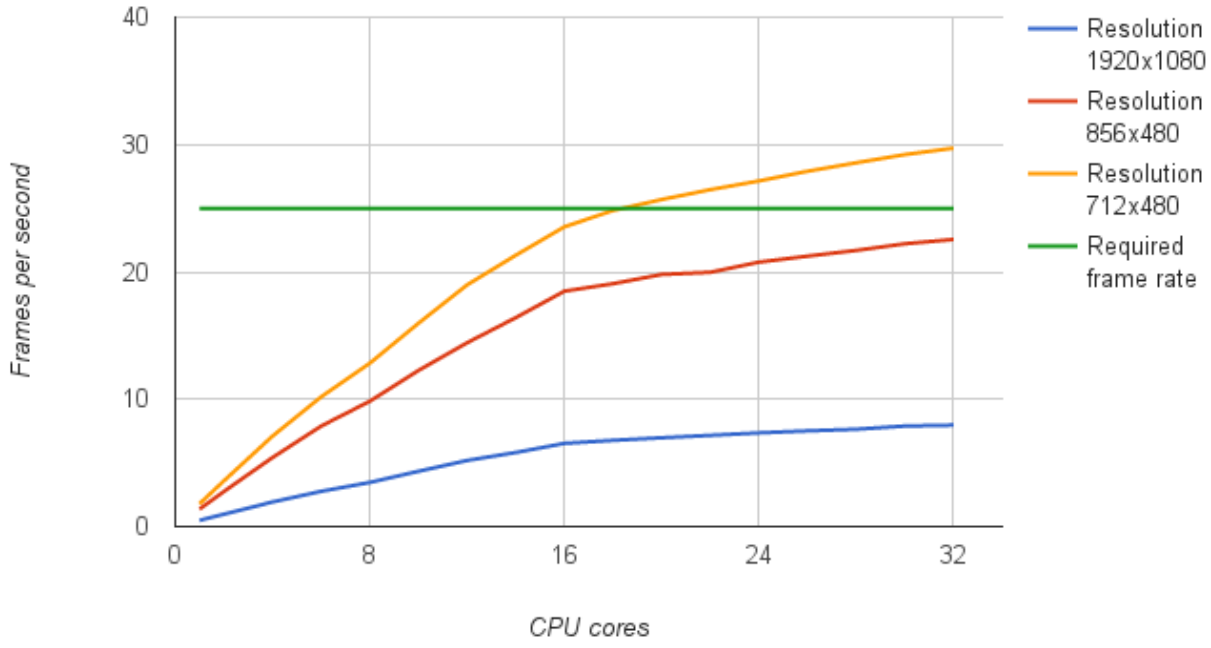



Figure 3.12: Localization performance in terms of FPS depending on the number of CPU cores and the resolution of the videos. The videos are the same as for the detection part. As the results show, the performance depends heavily on the resolution of the videos.

#### 3.4.4.2 Memory

We also tried to find out how the different parts of the system scale in terms of memory, and we investigated the influence of the index sizes on the performance. The memory usage for both parts is shown in figure 3.13 for detection and figure 3.14 for localization.

*Memory Performance for Localization:* For the localization, the memory usage behaves as expected (linear growth) and shows that the localization is scalable in terms of memory (as depicted in figure 3.14).

*Memory Performance for Detection:* For the detection, the memory usage behaves rather unusual after a certain number of used CPU cores (shown in figure 3.13). Therefore, a closer look into it was necessary. The results of this can be found in the figures 3.15 and 3.16. We tested different memory sizes used for the detection starting from 1GB up to 32GB. These tests showed that the available memory for the detection part does not influence the FPS performance (see figure 3.15). The Java memory scheduler uses as much memory as it can get, but it also performs well with only 1GB (shown in figure 3.16). This proves that the detection part is not dependent on memory, and therefore, memory is not a bottleneck for scaling the system, i.e., at least in the scope of the amount of data that we have available for our experiments [142].

Finally, we wanted to know if the size of a classification index influences the detection accuracy or processing performance. *Index Size Performance:* Figure 3.17 depicts the detection and processing performance as detection accuracy (F1 score) and FPS for three different index sizes. We expected that smaller indexes would lead to higher FPS throughput but with a loss of classification performance. Surprisingly, the experiment revealed that the tested index size does not have a significant influence on the FPS performance of the detection system. Of course, it is possible that an index with several hundred thousands frames will most probably lead to a lower

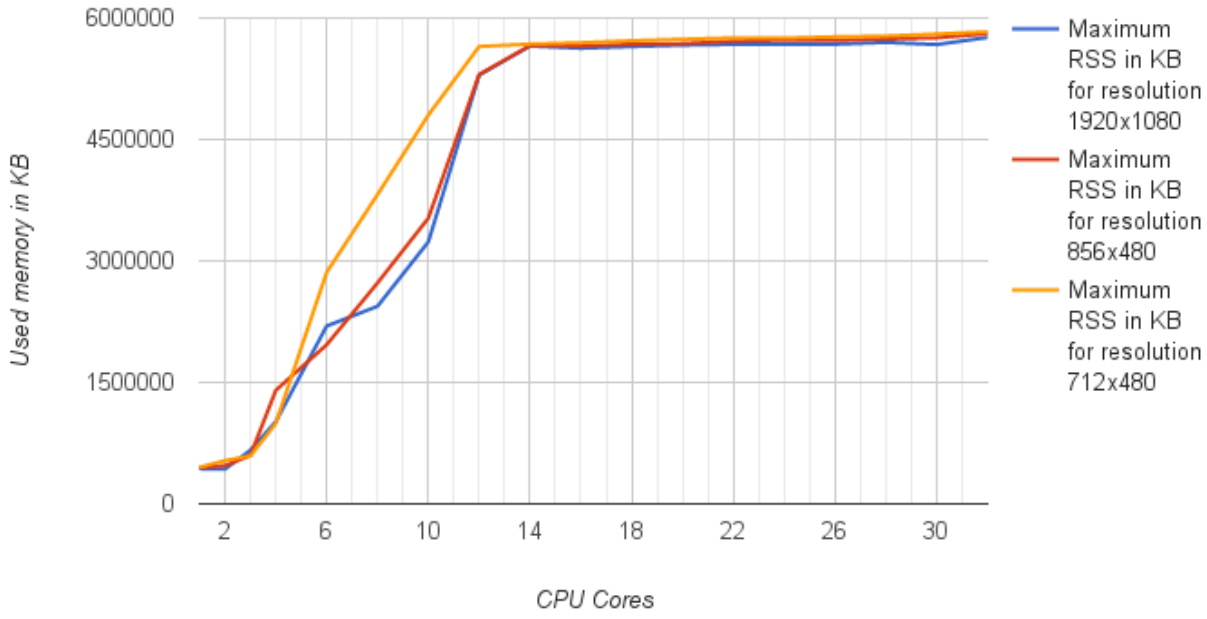



Figure 3.13: This chart shows the overall memory consumption for all three videos in the detection part. A maximum is reached at around 14 used CPU cores. Further investigation is needed to see if the detection part is scalable.

FPS output, but in the medical fields lack of training data is common and therefore this will not be a problem. Another positive aspect that we found in this experiment is that the classification performance does not decrease with smaller indexes tested (obviously a too small will). Even the opposite happened, for example, for the video with a resolution of $856 \times 480$ the F1 score increased slightly compared to the full training data. This is an indication that the detection part performs very well even with a smaller amount of training data. This finding is a very positive point for our use case, because of the constant lack of training data in the medical fields [142].

In this section, we presented our first evaluation of the EIR system and our findings based on this evaluation. We can report very good results for both the detection and localization accuracy and the processing performance. More detailed evaluations, based on the latest version of EIR, which includes GPU-offloading and is able to reach around 300 FPS, can be found in the papers [127, 143, 126, 142] and the following section.

## 3.5 Real-time Distribution of Multimedia Workloads in EIR

In order to process the videos from the GI tract using multiple filters in real-time, we need to parallelize the video analysis. In addition, to be able to scale to a massive scale cancer screening of large parts of the population every year, the system must be distributed [126, 127].

Major limitations of the current state-of-the-art is the lack of support for (i) efficient execution of large workloads like the VCE on elastic heterogeneous resources in general and (ii) delivery of results in real-time. To achieve efficient processing and real-time capabilities, data must be efficiently managed and processed, i.e., the current trend where everything is pushed to the cloud does not necessarily work alone due to huge data sets, varying resource availability, privacy issues and real-time constraints.

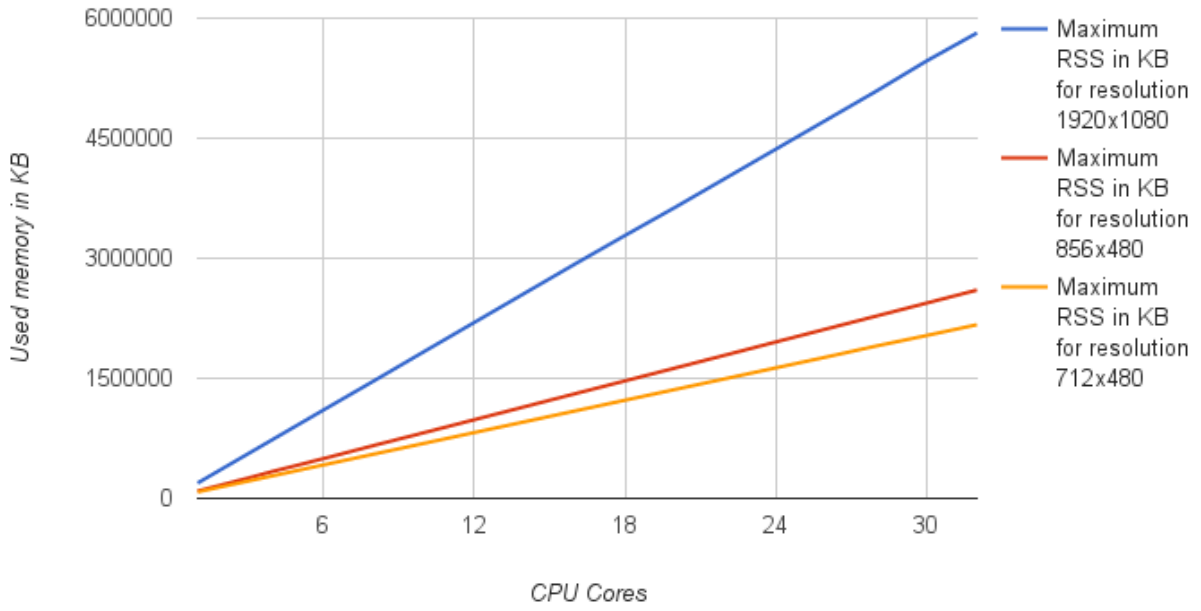



Figure 3.14: This chart shows the overall memory consumption for all three videos in the localization part. This shows us that the localization part scales well in terms of memory.

In this respect, distributed and parallel processing has been important and popular research area for a long time, and in a never-ending quest for solving increasingly complex tasks, software developers have consistently pushed the boundaries for computational demand. As the physical limitations of pushing the clock frequency on a single-core CPU became apparent, new hardware solutions evolved. The focus moved to increasing the number of cores. This effectively forced the programmers to parallelize their applications to increase performance.

After we proved that the EIR system achieves good results in terms of detection and localization accuracy, we started to improve the processing performance so that it can be used in real-time examinations. For this, we utilized heterogeneous resources such as multiple CPUs and GPUs [142, 126, 127]. We also performed some distributed processing experiments on Amazon mechanical cloud [142] that showed us that the performance gain is not as high as expected and distribution on multiple GPUs and CPUs is efficient enough. Moreover, we implemented a version of EIR that utilizes device lending (virtual lending of resources like GPUs from other machines via PCI Express) to improve the performance even more [74, 126]. The EIR system has been implemented completely and showcased in several demo session [145, 126] to show its real-time capabilities. This section gives a summarized overview of the research conducted in area of processing performance, a description of our GPU acceleration and how we utilized device sharing using device lending.

### 3.5.1 Distribution and Offloading of Multimedia Workloads

With parallelization, modern distributed computing systems often provide the required processing power for large scale data processing. However, their increasing complexity is a challenge. Existing and future topologies exhibit a range of capabilities where computing nodes consist of multiple heterogeneous processing engines including multi-core CPUs, digital signal processors (DSPs), field-programmable gate array (FPGAs) and GPUs. To further complicate the

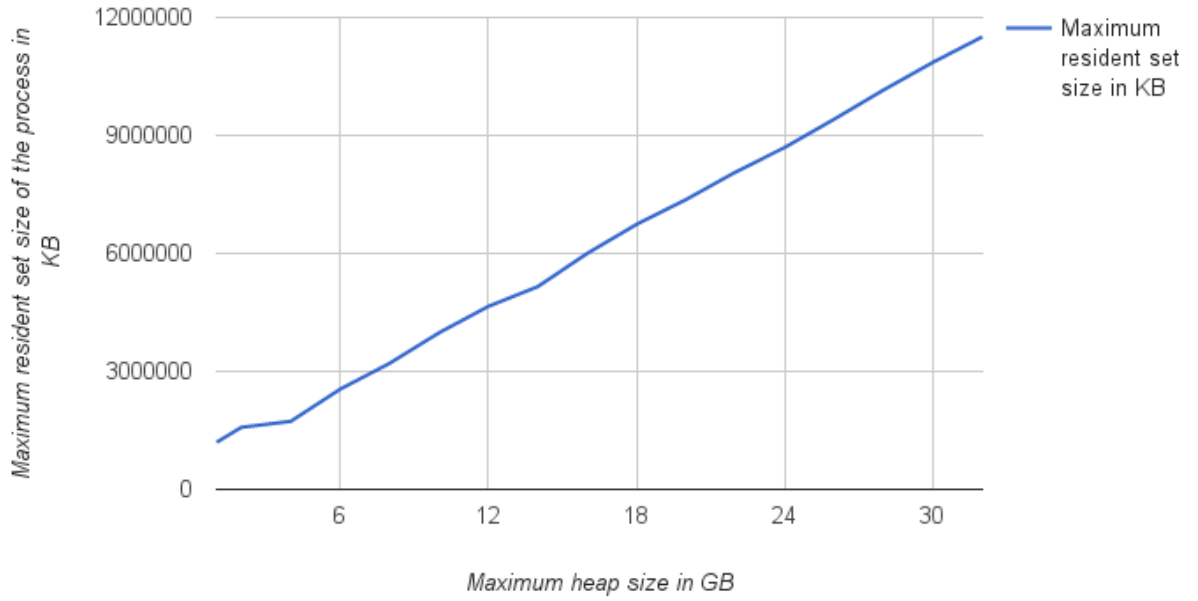



Figure 3.15: The analysis of the memory consumption of the detection part showed us that the Java garbage collector always uses the complete memory that it can get. It is automatically set to around 6GB on our system.

picture, these devices often vary their behavior between generations [120]. Furthermore, these nodes are connected by a variety of networks, from wireless networks to high-speed interconnects. Communicating via a network adds latency and can introduce delays detrimental to the progress of the application. Moreover, parallel applications on shared-memory architectures need to synchronize threads of execution when accessing shared data. It becomes necessary to distribute data and synchronize state between nodes in the topology. Fault-tolerance is required to handle situations where nodes in the network become unavailable, and these are just a few of the obstacles that must be overcome. Thus, the increased complexity results in a lack of portability of the existing scheduling designs, and moving from a *sequential* mode of operation to parallel execution is significantly *more demanding* for developers [107, 119].

A first wave of building blocks for big-data systems like, for instance, *Apache's* Hadoop, Hive, Spark, and Storm, *Google's* MapReduce, GFS, Chubby, and BigTable, *Microsoft's* Dryad, Azure and Naiad, and *Facebook's* Zookeeper, and Cassandra have already been developed. These are isolated building blocks related to handling high data volumes, and not complete big data systems [35].

Other novel solutions are expected in the years to come since a substantial subset of international computer science has redefined their research agenda under the big data umbrella. Prominent examples include MIT's bigdata@CSAIL, Berkeley's AMPLab, and Ireland's INSIGHT Center, which also work in the intersection with medicine [36, 107].

All previous mentioned frameworks are limited to certain application domains, i.e., *batch processing*, and some later approaches *stream processing*. Processing and analyzing thousands of videos from, for example, VCEs in real-time, potentially using a long pipeline of filters to search for multiple disease scenarios, gives other requirements, and existing frameworks cannot be used out-of-the-box. We need greater flexibility to extend the processing pipeline, support for iterations, deadlines for immediate feedback and the ability to express arbitrary processing

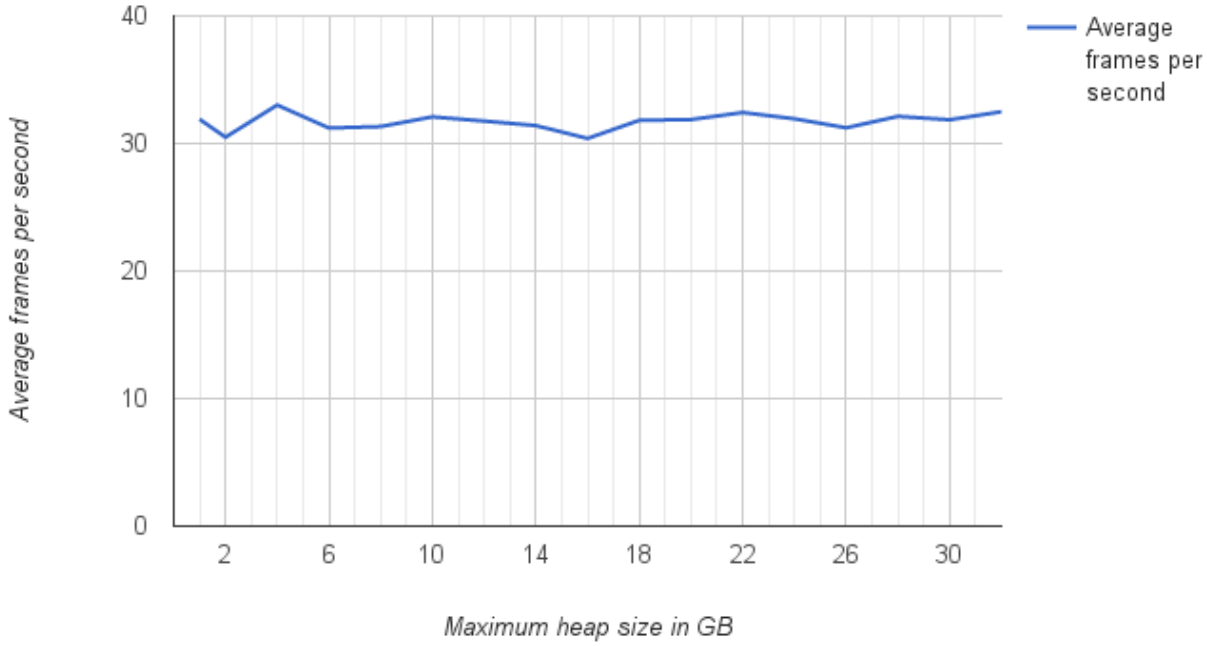



Figure 3.16: This experiment showed that the available memory for the detection part does not influence the FPS performance. The Java memory scheduler takes always the whole memory that it can get but it also works perfectly with only 1GB. This is a proof that the detection part is not dependent on memory and therefore memory is not a bottleneck for scaling the system.

graphs [36]. Thus, traditional batch processing frameworks do not commonly integrate knowledge of deadlines into the run-time itself. While support for iterations exists, it is not considered in the parallelization stage, and it is rather solved by iterative execution of the workload, which might introduce artificial barriers between iterations, potentially slowing down faster workers [107]. In the beginning of this work, we started to investigate different distributed frameworks regarding their possibilities and to determine the state-of-the-art. Therefore, we investigated Hadoop [190], Apache Spark [161], Cuda [115] and Apache Storm [164]. Reviewing of the literature and testing existing frameworks showed us that the Apache Storm framework is the best solution for our requirements and that it might could be used as a high level distribution framework for P2G kernels (P2G is a distributed framework specifically designed for multimedia workloads [35]) that we wanted to use within this work.

To be able to schedule multimedia workloads efficiently, more dimensions than just the time (e.g., workloads and computational attributes) have to be taken into account. This introduces new concepts for scheduling as a combination of taking a computational tasks nature, in term of if it is time-, CPU-, I/O-, memory- and cache-bound, into consideration.

CPU bound means the rate at which process progresses is limited by the speed of the CPU. A task that performs calculations on a small set of numbers, for example multiplying small matrices, is likely to be CPU bound [197]. In P2G, because we have fetch statements, we know how much data is being requested for the computation done by every kernel. Because of this, it is possible to look at the ration between instructions being generated vs. the data fetched from fields. It might not be accurate every time, but it can give an indication on whether the kernel is CPU bound or not.

I/O bound means the rate at which a process progresses is limited by the speed of the I/O


Figure 3.17: This chart shows how the amount of training data influences the performance of the detection subsystem in terms of detection accuracy and FPS 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). The chart shows that there is no significant difference for the detection performance and the FPS. The smaller indexes can achieve even a better F1 score for the video with a resolution of $856 \times 480$ [142].

subsystem. A task that processes data from disk, for example, counting the number of lines in a file is likely to be I/O bound [8]. P2G does not support this functionality yet, but source/sink kernels [125] that handle I/O are planed for the future. This is important because of distribution where kernels can be migrated to computers that do not have the correct files. Furthermore, having ”special I/O fields”, e.g., a kernel fetching from an I/O field would implicitly become a source kernel and a kernel storing would implicitly become a sink kernel might be a promising idea.

Memory bound means the rate at which a process progresses is limited by the amount memory available and the speed of that memory access. A task that processes large amounts of in memory data, for example multiplying large matrices, is likely to be memory bound [174]. In general, this is related to CPU bound. In P2G, this can be statically assessed, e.g., generate Low Level Virtual Machine (LLVM) Intermediate Representation (IR) [78] with clang reads in generated code and use the code to look up fetches which provides the data type. These data type sizes can then be aggregated for every fetch and compared to the number of compute operations found in the LLVM IR.

Cache bound means the rate at which a process progress is limited by the amount and speed of the cache available. A task that simply processes more data than fits in the cache will be cache bound. I/O bound would be slower than memory bound would be slower than cache bound would be slower than CPU bound [41]. The solution to being I/O bound is not necessarily to get more memory. In some situations, the access algorithm could be designed around the I/O, memory or cache limitations like cache oblivious algorithms [151]. As an optimal solution, it might be possible to implement cache oblivious processing with P2G kernels, but at first one

should look into basic loop/block tiling [194, 122]. This can be achieved by looping over a (sub-)graph and not process the complete dataset at once, but instead smaller chunks of it.

As an alternative to this, one can basically compute anything on the GPU [121]. The most important classification here is whether a problem is task parallel or data parallel [76]. The first one refers, roughly speaking, to problems where several threads are working on their own tasks, more or less independently [82]. The second one refers to problems where many threads are all doing the same - but on different parts of the data [63]. The latter is the kind of problem that GPUs are good at: They have many cores, and all the cores perform the same task, but operate on different parts of the input data. An example problem for that would be "simple math but with huge amount of data". Although this may sound like a perfectly data parallel problem, and thus like it was well-suited for a GPU, there is another aspect to consider: GPUs are ridiculously fast in terms of theoretical computational power (Floating Point Operations Per Second, FLOPS), but they are often throttled down by the memory bandwidth [155]. This leads to another classification of problems. Namely, whether problems are memory bound or compute bound, which we already mentioned at the begin from a more general point of view.

The first one refers to problems where the number of instructions that are performed for each data element is low. For example, a parallel vector addition: First, we have to read two data elements, then perform a single addition, and then write the summary into the result vector. This will not be faster if performed on the GPU, because the single addition does not compensate for the efforts of reading/writing the memory [120]. The P2G kernels are created to distinguish task and data parallel operations. Index expressions are data parallel and independent kernels are always task parallel. Like using a GPU, P2G could identify "a long string" of data parallel kernels. Agglomerating these kernels into one larger, or longer running, kernel can hide the penalty associated with transferring data over the PCI express bus. It is also possible to analyze the generated LLVM IR and identify loops and excessive branching in kernel's native code. Branching control structures are not ideal for GPUs [62].

The second term, compute bound, refers to problems where the number of instructions is high compared to the number of memory reads/writes. For example, a matrix multiplication: The number of instructions will be $O(n^3)$ where $n$ denotes the size of the matrix. In this case, one can expect that the GPU will outperform a CPU at a certain matrix size. Another example could be when many complex trigonometric computations (like sine/cosine) are performed on *few* data elements [120]. P2G and the new LLVM IR that is used should be able to easily identify sine operations (at least that is the impression, because LLVM has what it calls internal intrinsics, i.e., common functions such as memcpy, malloc, free, and also cos and sin), which are efficiently executed on the GPU [75]. As a rule of thumb it can be assumed that reading/writing one data element from the *main* GPU memory has a latency of around $500$ instructions. Therefore, another key point for the performance of GPUs is data locality. If it is needed to read or write data, and in most cases this is true, then should be made sure that the data is kept as close as possible to the GPU cores. GPUs thus have certain memory areas (referred to as local memory or shared memory) that usually is only a few KB in size, but particularly efficient for data that is about to be involved in a computation [146].

To emphasize this, GPU programming is an art that is only remotely related to parallel programming on the CPU. Functionalities like threads in Java, with all the concurrency infrastructure like ThreadPoolExecutors, ForkJoinPools, etc., give the impression that tasks just have

to be split up and distributed among several processors [45]. Unfortunately, it is not that easy. On the GPU, one can encounter challenges on a much lower level, namely, occupancy, register and shared memory pressure and memory coalescing. However, if a data-parallel, compute-bound problem has to be solved, the GPU is most probably the best solution [81]. In P2G, we are most likely able to identify these exact properties by examining kernels. In addition, we can also estimate close values for calculating the GPU occupancy. This is done manually by programmers today using an excel sheets.

Unfortunately, due to its huge scope and complexity, the development of the P2G framework got delayed due to a change of the code basis to LLVM, and we discovered that timing within scheduling of multimedia workloads was an already well researched field [188, 94, 67, 152]. P2Gs current state is rather undeveloped and it can not really be used to conduct tests on multimedia workloads. Essential parts like file reading and writing and memory management are missing. To make P2G working as intended and that it can be tested in a real scenario will need some more years of engineering.

Taking all these aspects into the scheduler makes it a very complex and hard to solve task. Some literature research showed us that hyphergraphs are a promising tool to solve these complex problems, which introduces a lot of new challenges and research questions that are interesting to address. Nevertheless, because of the current state of P2G all of them require a lot of engineering and programming time to be able to be tested and are therefore out of scope for this thesis. Instead, we decided to start with exploring the possibilities of using GPUs to increase the processing performance of parts of the of the EIR system. This will give a first idea about the problems and challenges of our workloads and lay the basis for future work.

### 3.5.2 GPU-acceleration

To improve the processing performance of the EIR system, we implemented some of the compute-intensive parts using GPUs and Cuda [115]. Cuda is the standard framework that is used to program Nvidia graphic cards. We decided to pipeline parts of the system architecture as heterogeneous processing subsystem. An overview of the implementation can be found in figure 3.18.

The main processing applications (indexer and classifier) interact with a modular image processing subsystem. Both parts are implemented in Java via direct calls of the API. At the moment, the GPU-accelerated processing supports a number of features (JCD, which includes FCTH and CEDD, and Tamura). We decided for these as starting points because they are the best performing ones for our use case. Specifically, we implemented the feature extraction, color space conversion, image resizing and pre-filtering [142, 127]. To be able to handle multiple image processing and feature extraction requests at the same time, the image processing subsystem uses a multi-threading architecture. The GPU implementation of EIR is transparently accessible from native Java code through a GPU CLib wrapper.

The Java Native Access (JNA) API [89] is needed to directly access the GPU CLib API from the image processing subsystem. The GPU CLib is a system shared library and implemented in C++. The main task of the GPU CLib is to maintain the connection and handle data streams with the stand-alone Cuda-enabled processing server. To achieve maximum data transfer performance and reduce excess data copy actions, shared memory is used. To send requests and receive status responses from the Cuda server, local Unix sockets are used, which are easy to

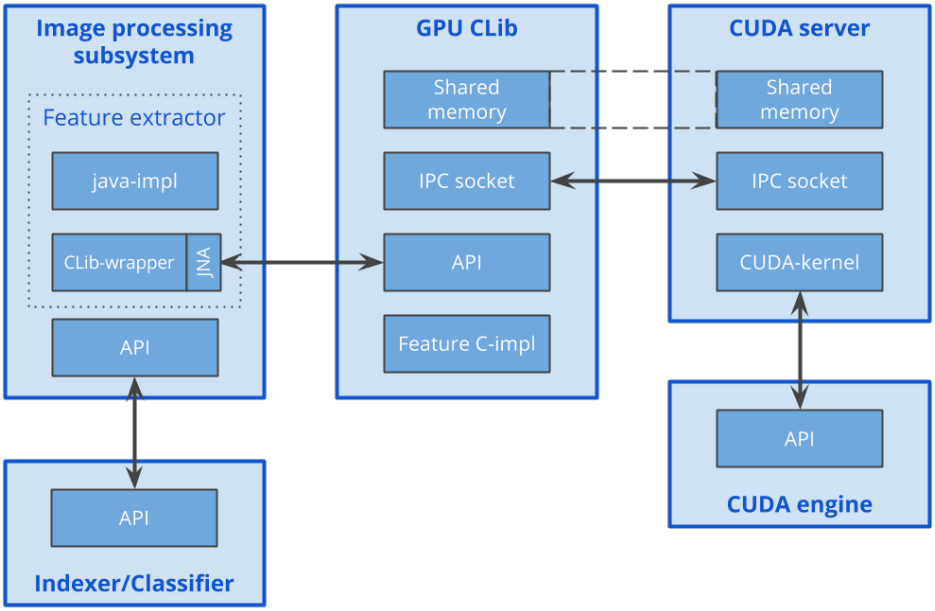



Figure 3.18: The main processing application consists of the indexing and classification parts and uses the GPU-accelerated image processing subsystem to increase the processing performance. The image processing subsystem provides feature extraction and image filtering algorithms for the pipeline. Compute-intensive procedures are executed by a stand-alone Cuda-enabled processing server. The interaction between the different architectures is performed via a GPU CLib shared library which is responsible for maintaining connections and handling data streams with the Cuda-server [127].

utilize for communication between processes executing on same host operating systems. The Cuda server is implemented using C++ and uses Cuda SDK to perform the processing tasks on the GPU. The server and all supporting components are built with distributed processing in mind, and can easily be extended with multiple Cuda servers running locally or on several distributed nodes [127]. The processing server can easily be extended with new extractors for different features and advanced image processing algorithms if needed, which can benefit form the provided utilization of multi-core CPU and GPU resources. As an example of a feature extraction, the extractor for the FCTH feature is depicted in figure 3.19. One can see that for the image features, all pixel-related calculations are executed on the GPU. This also includes the processing of multi-threaded shape detector and fuzzy logic algorithms which are image processing steps that are needed for FCTH. The heterogeneous processing subsystem also provides input and intermediate data transparent caching services. These services help to reduce the CPU-GPU bandwidth usage and avoids excessive data copying during the processing steps within the image processing part [142, 127].

#### 3.5.2.1 Performance Evaluation

Utilizing the processing power of GPUs in the EIR system increased the performance greatly compared to the CPU only version. To evaluate and compare the performance gain, we per-

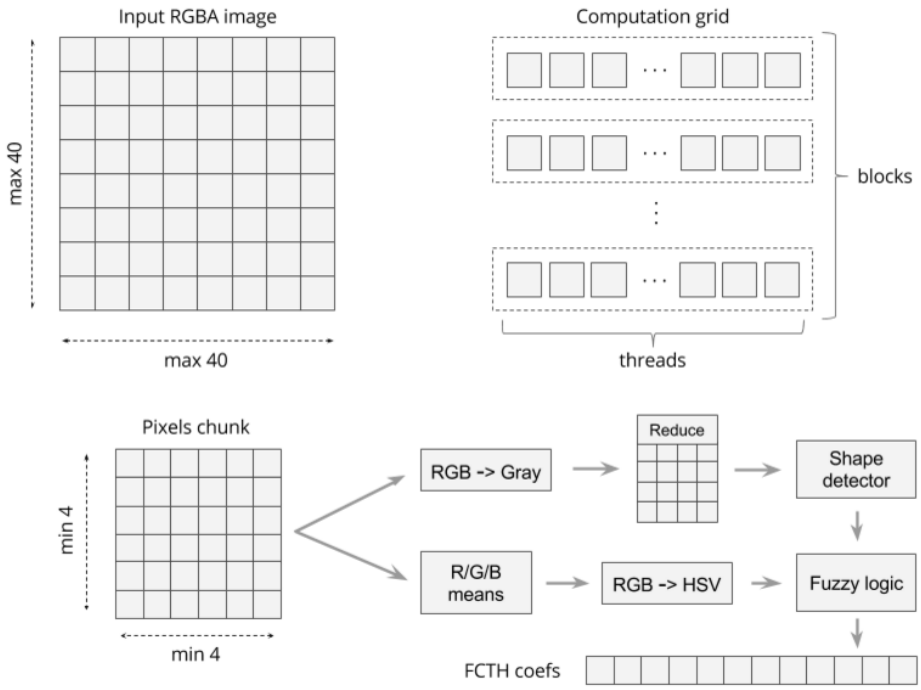



Figure 3.19: The figure shows an example of our FCTH feature implementation using the GPU extension of the EIR system. The input image is split into a number of non-overlapping blocks that can be distributed. Each of the blocks is processed by two GPU-threads. The main processing steps include color space conversion, size reduction, shapes detection and fuzzy logic computations [127].

formed experiments on the ASU Mayo dataset. We used three videos with different resolution to measure the exact increase of the performance. The discussed experiments are presented in more detail in the paper [127]. The used resolutions are full HD ($1920 \times 1080$), WVGA1 ($856 \times 480$), WVGA2 ($712 \times 480$) and CIF ($384 \times 288$) [142]. The results are presented in figures 3.20, 3.21, 3.22 and 3.23. For these experiments, we decided to use $30$ FPS or $33.3$ milliseconds per frame as target processing time for live, real-time examinations [127]. The hardware used for the experiments was a standard desktop computer, Intel Core i7 3.20GHz CPU, 8GB RAM and a GeForce GTX 460 GPU. The basic and improved EIR systems are compared using the same Java source code. In the figures, the basic system's results are labeled as Java, the improved system's results with disabled GPU-acceleration are labelled as C and the improved system's run in the heterogeneous mode with enabled GPU-acceleration is labeled as GPU [142, 127].

We can observe in figure 3.20 that the basic architecture can process full HD frames using all $8$ available CPU cores and up to $4$GB of memory only at $6.5$ FPS for Java and $13.8$ FPS for the C implementations. For the smaller frame sizes, real-time requirements are reached using a maximum of $4$ CPU cores and a maximum of $4$GB of memory. The maximum FPS that were achieved are $49$ FPS, $51$ FPS and $66$ FPS for WVGA1, WVGA2 and CIF as figure 3.21 and figure 3.23 show. The new GPU-enabled architecture easily process full HD frames using only $4$ CPU cores and $5$GB of memory with $32.6$ms frame processing time (see figure 3.20 and 3.22). The maximum frame rate for full HD video frames was $36$ FPS using all $8$ CPU

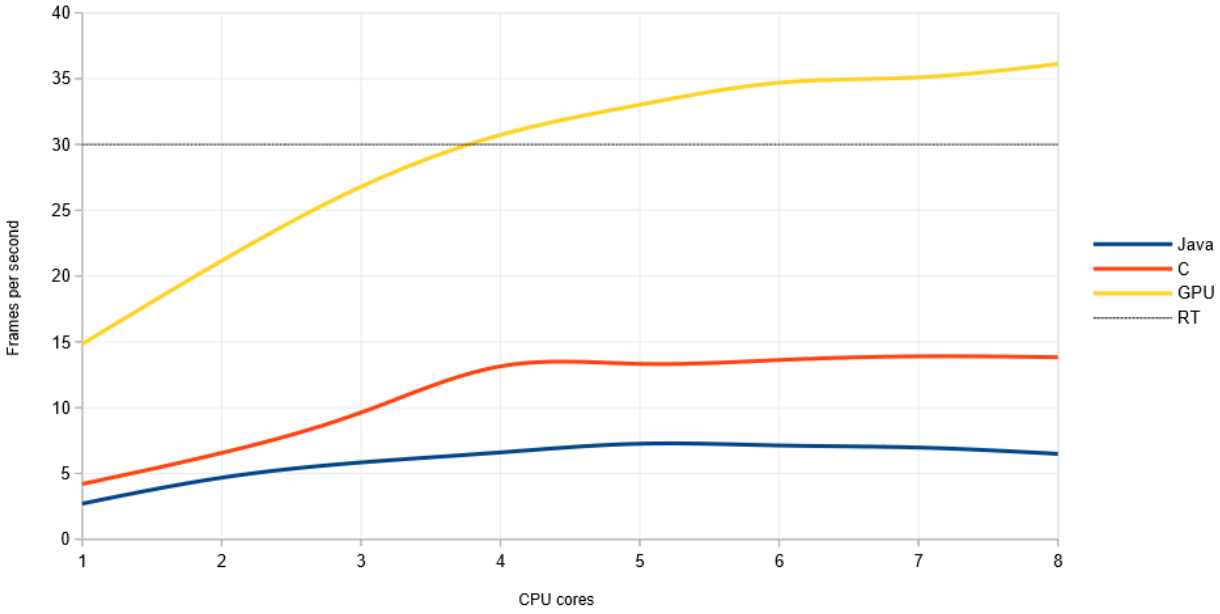



Figure 3.20: This image show the performance of the improved EIR system for full HD frames. It reaches real-time performance (RT line) with 30 FPS for full HD ($1920 \times 1080$) videos on conventional desktop PC using only 4 CPU cores and 5 Gb of memory. The maximum frame rate is around 36 FPS using 8 CPU cores. The Java and C implementations cannot reach real-time performance on the used hardware [127].

cores. The smaller frame sizes reached real-time with only 1 CPU core and 4.5GB of memory. The maximum frame rate achieved by the EIR system where around 200 FPS (see figure 3.21 and figure 3.23). The tested hardware with the basic EIR system is not able to achieve real-time performance for full HD videos even using all available CPU cores (but it works for the low resolution videos). The improved EIR system reaches real-time performance for full HD videos using 4 CPU cores and one rather outdated GPU. The smaller frames can be processed using very little resources (one CPU core plus the GPU). The memory size is never a limiting factor (which could be a problem with GPU implementations), and the improved EIR system can be deployed even using a single desktop PCs with a standard GPU as an accelerator. We underlined this finding with further experiments which are described in more detail in the papers [142, 126, 127] where we achieved FPS rates up to 300 FPS for HD resolutions.

### 3.5.3 Device Lending

To explore further possibilities for improving the performance of the EIR system, we tested it using device lending [74]. Device lending is a transparent cross-machine device sharing system without any need to implement application-specific distribution mechanisms that allows to share resources like GPUs between different machines. For applications run on hardware that use device lending, the remote I/O resource appears local and does not have to be addressed in a specific way. This makes it very easy to use and interesting for systems that for example benefit from GPU acceleration like our improved EIR version. Device lending is implemented on top of native PCI Express which enables low-latency and high-throughput and with PCI Express as the most widely used I/O bus today, it is also already present in all modern computers [74].

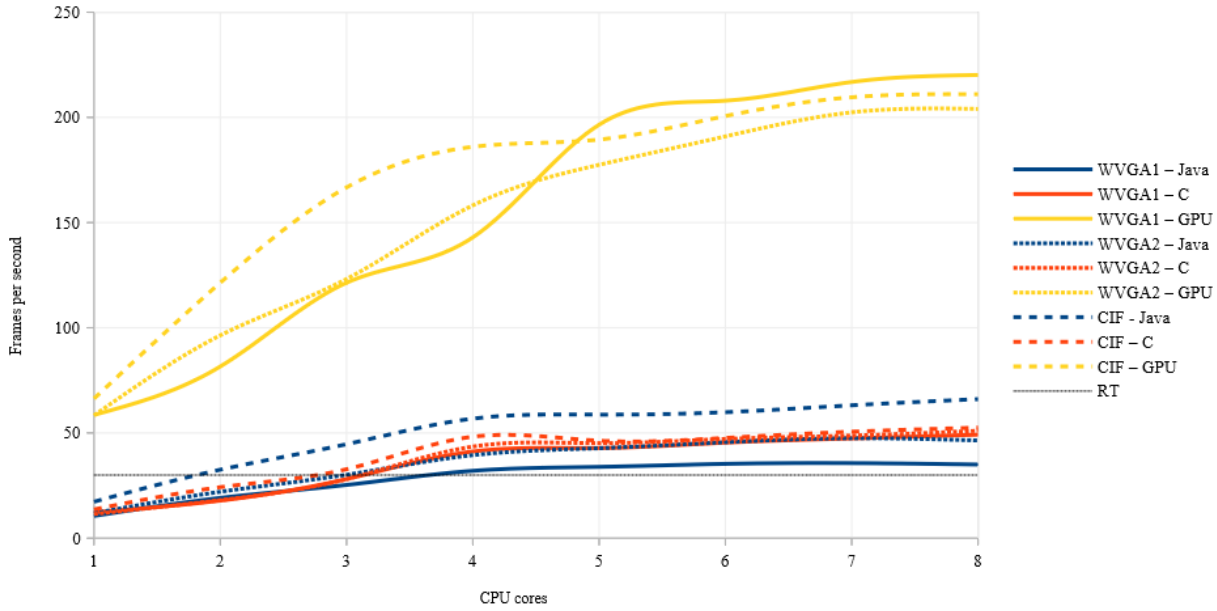



Figure 3.21: This figure shows the performance of the EIR system for non HD frames. The videos WVGA1 ($856 \times 480$), WVGA2 ($712 \times 480$) and CIF ($384 \times 288$) can be processed in real-time by the improved EIR system using only $1$ CPU core. The maximum frame processing rate reaches more than $200$ FPS [127].

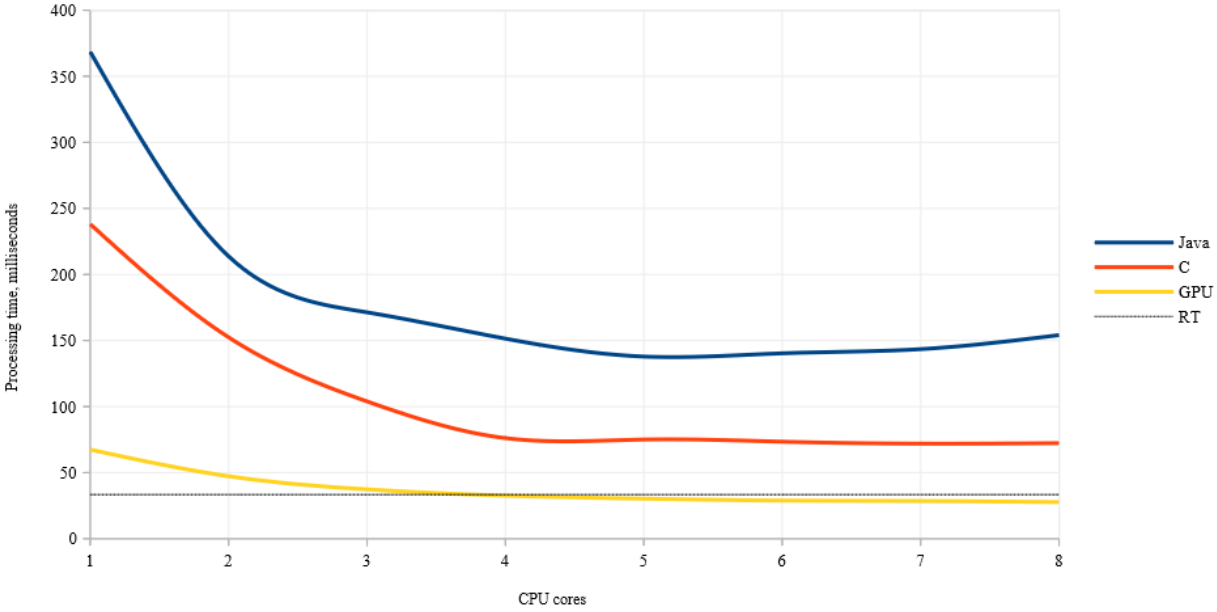



Figure 3.22: The processing time decreases marginally with an increasing number of used CPU cores for a single full HD frame. This is due to the CPU-parallel implementation of feature comparison and search algorithms which are not as compute intensive as the feature extraction processes. Java and C implementations reach the required frame processing time with $4$ CPU cores (hyper-threading cannot handle CPU intensive calculations efficiently for all $8$ possible which are $4$ real and $4$ virtual cores on the used system) [127].

For our use case, this is interesting because it can be beneficial to have one main frame that can lend the devices to different computers based on the requirements in a hospital scenario where resources and space in the examination rooms are often spare [74].

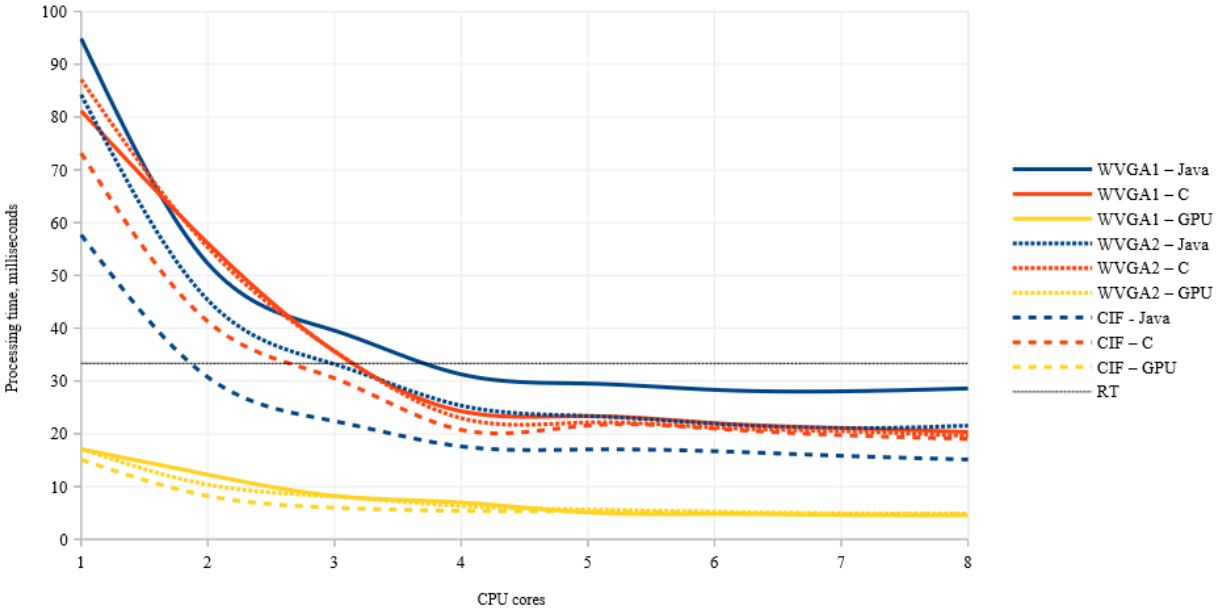



Figure 3.23: Using EIR with GPU support for processing smaller frame sizes results in a processing time far below the real-time margin. The minimum is reached with 5 milliseconds. This is a prove for the high system performance and ability to be extended by additional features or to process several video streams at the same time [127].

| Device | Type | E1 | E2 | E3 | E4 |
|---|---|---|---|---|---|
| CPU | Intel Core i7-4820K 3.70GHz | * | * | * | * |
| GPU1 | Nvidia Tesla K40c | * | * | * | * |
| GPU2 | Nvidia Quadro K2200 | | * | * | * |
| GPU3 | Nvidia GeForce GTX 750 | | | * | * |
| GPU4 | Nvidia Tesla K40c | | | | * |
| RAM | 16GB Quad Channel DDR3 | * | * | * | * |

Table 3.9: This table shows the used hardware and the configurations for the different conducted experiments. GPU1 to GPU3 are local GPUs and GPU4 is lent via device lending [126].

#### 3.5.3.1 Performance Evaluation

To evaluate the performance increase when using device lending for our system, we performed four different experiments [126]. An overview of the performed experiments can be found in table 3.9. We used the same CPU (Intel Core i7) and memory (16GB) for all experiments, and we used two computers. One of them is used to perform the tests and one is only used for lending the GPU [74].

The experiments are labeled from E1 to E4. E1 uses one local GPU, E2 uses two local GPUs and E3 used three local GPUs and E4, borrows one GPU from the second computer in addition to three local GPUs [74, 126]. In the experiments, we performed the polyp detection task with our EIR system and real-time feedback as an overlay on the video for up to 16 parallel video streams. All video streams are full HD ($1920 \times 1080$). We measured the performance for the complete pipeline from capturing the video until showing the output on the screen. The results of the experiment are presented in the figures 3.24 and 3.25 [74, 126].

Figure 3.24 shows the performance for all streams streamed at the same time. For a max-

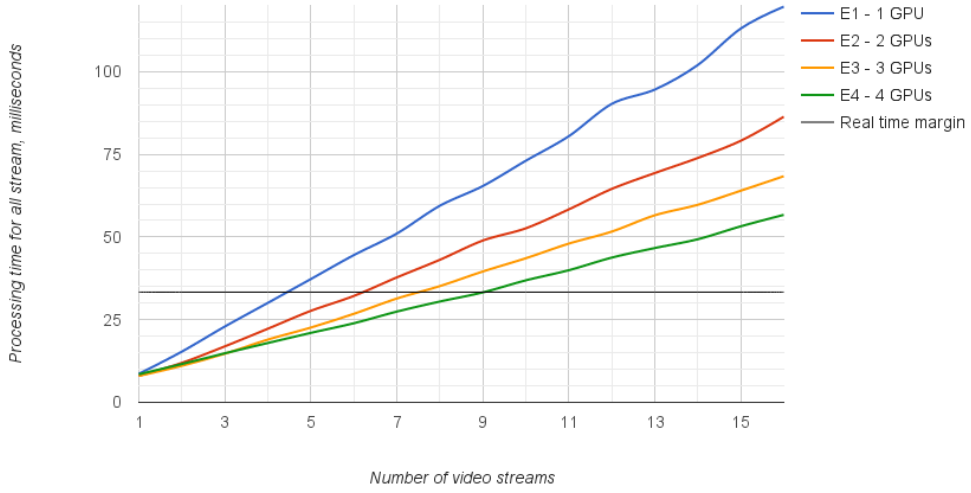



Figure 3.24: Frame processing time for several full HD streams in parallel using the different experimental setups for GPU acceleration (table 3.9) [126].

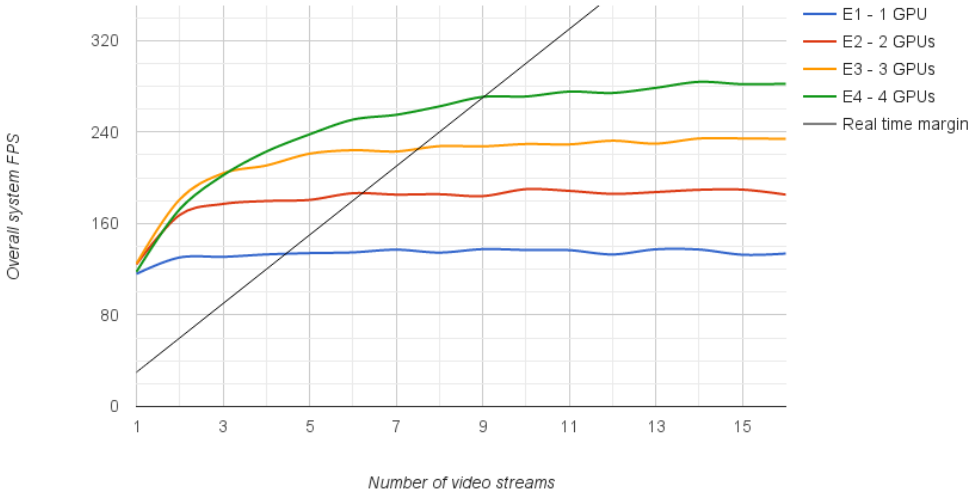



Figure 3.25: The overall system performance of multiple video streams in parallel for all experimental setups using GPU acceleration [126].

imum of seven parallel full HD streams, the three local GPUs are fast enough. If more than seven parallel streams are used, GPU lending is required. One can see that the more parallel streams are processed the better is the performance gain from the device lending technology. This is because the overhead for transferring data that hinders device lending using its full potential. Figure 3.25 shows the maximum system performance which reveals that device lending can indeed improve the systems performance. The maximum FPS reachable using four GPUs at the same time is 30 FPS for nine full HD streams streamed in parallel. This would be around 270 FPS for a single video stream. Another insight is that the device lent GPU does not increase the performance for a small number of videos but for more than five videos the increase gets noticeable. This is an indicator that device lending is not helpful for a small amount of data but that it can increase the performance largely for larger amount of data [126].

## 3.6 Proof-of-concept for Multi-disease Classification

In section 2.2.4, we claimed that one major difference between EIR and related approaches is that EIR easily can be extended to detect other endoscopic findings (abnormalities, diseases, anatomic landmark or other relevant events during the examination of a patient).

To proof that EIR is able to perform multi-class classification for diseases aside from polyps, we developed an improved prototype for the detection part, and additionally, a deep-learning-based approach to be able to compare it with a state-of-the-art method. Both approaches are tested on a dataset collected from the Bærum Hospital in Norway, which is one of our collaborators. The annotated data that we had to evaluate multi-class classification was rather limited and therefore it is important to point out that these results are preliminary. A more detailed evaluation is recommended when more annotated data is available (getting annotated data is an ongoing process and can take more than one year because of the workload for the doctors).

### 3.6.1 Multiclass-EIR

The EIR system is designed in a way so that multi-class classification can easily be added with minor adjustments. Figure 3.26 gives a detailed overview about how the multi-class version of EIR, called Multiclass-EIR, works. The basic search-based classification part of EIR is used to create a classifier for each disease that we want to classify. The difference to the two-class version 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 [16]. It is important to point out that for this step also other classification algorithms can be used. We decided for the random forest approach because it is fast and achieves at the same time good results [180].

The random forest classifier creates automatically a random forest of classification trees. The created classifier consists of different randomly generated decision trees. A decision tree can be seen as a classifier, which basically performs decision-based classification on the given data. To determine the final class, the classifier combines all decisions trees into a final decision (which is similar to late fusion described in section 2.3.5.2). The advantage of the random forest algorithm is that the training of the classifier is very fast because the classification steps can be parallelized, which is possible due to the fact that each tree is processed separately. Further, the random forest classifier 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 the training time has a linear increase with the number of trees, which can lead to a longer training time if many trees are used at the same time. However, this is not a problem for our use case since training time is not critical. Our implementation of the random forest classifier is using the version provided by the Weka machine learning library [54].

### 3.6.2 Deep-EIR

The Deep-EIR version of the detection part of EIR is based on deep learning. In particular, we trained a model based on the Inception v3 architecture [167], which is a deep learning architecture designed for image classification, using the ImageNet dataset [29]. The ImageNet

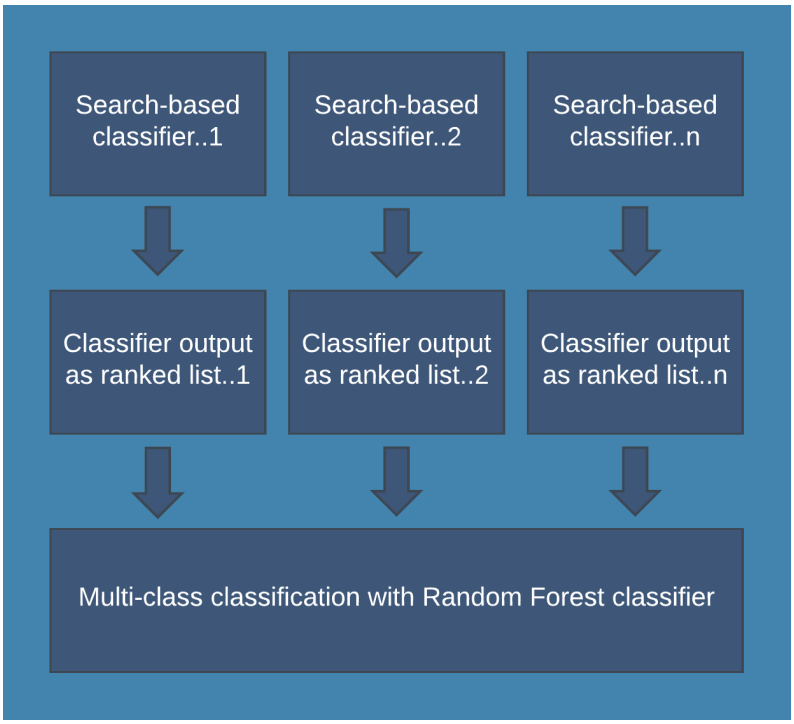



Figure 3.26: Detailed steps for the multi-class detection part of the EIR system. Several search-based classifiers are used for different classes, which are combined using an additional classification method.

dataset is a large dataset of annotated images containing $14,197,122$ images and $1,000$ object categories. From the Inception v3 model, we removed the last layer and retrained it with our medical image classes. Figure 3.27 gives a detailed overview of the complete pipeline for the neural-network-based implementation of the detection.

The Inception v3 achieves good results regarding single frame classification and efficient resource and parameter needs. The top one result error is around $21\%$, and the top five error is around $6\%$ with less than $25$ million parameters. The training of the Inception v3 network is performed from scratch using Goolge Tensorflow [1] and takes around three weeks on a single computer with GPU support. Tensorflow is an open source framework that allows all kind of numerical computing 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 [31]. Basically, we remove the final layer from the model 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 from Tensorflow[3]. 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.

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

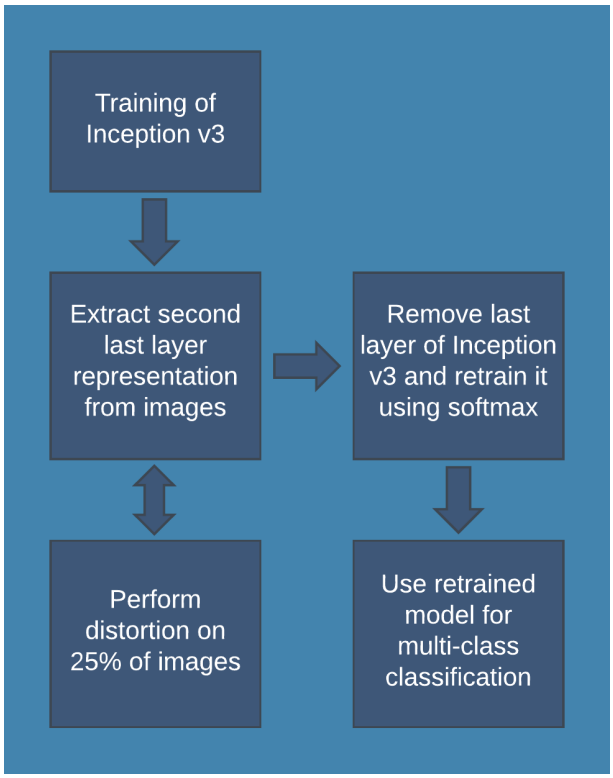


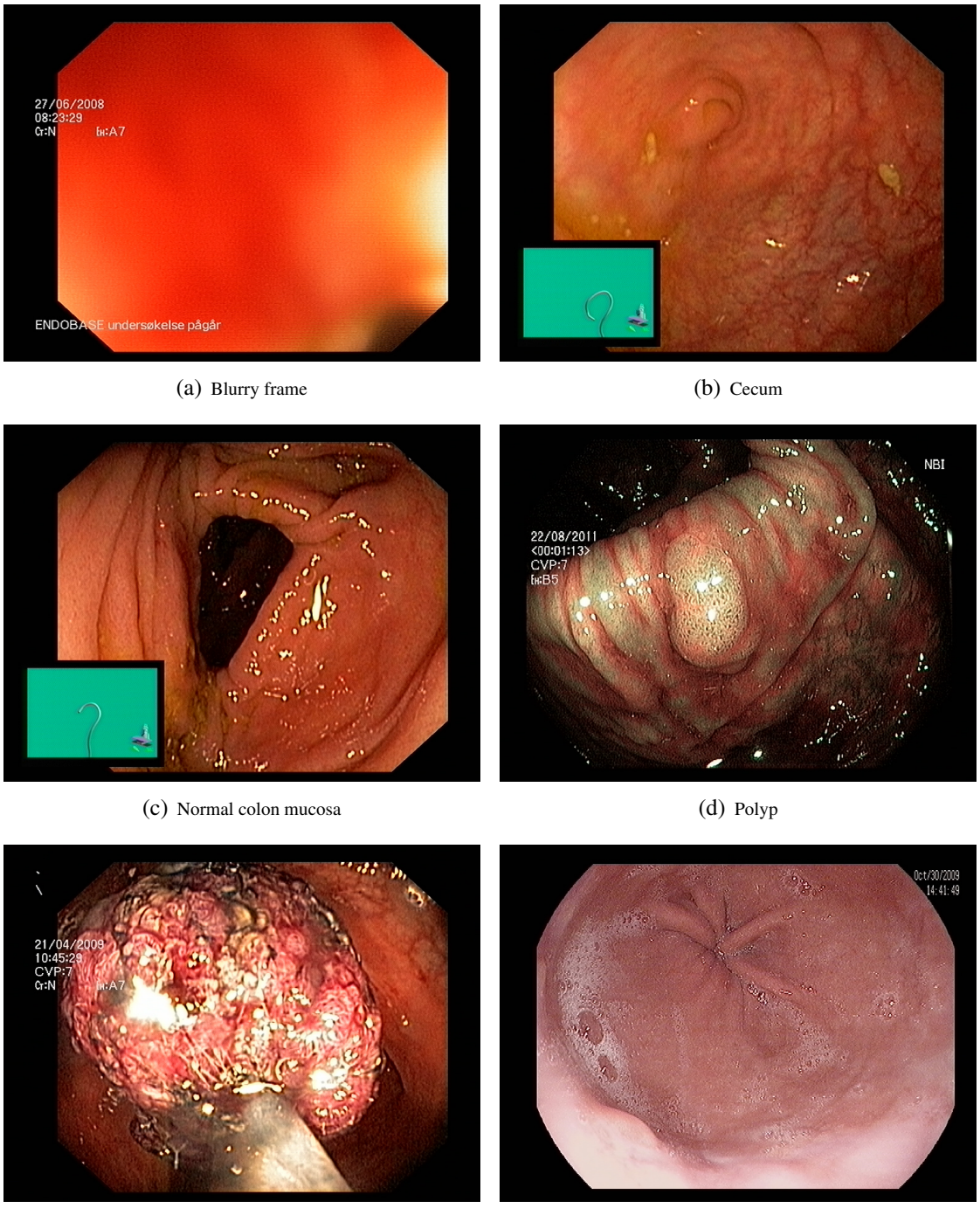

Figure 3.27: Detailed steps for the neural network (deep learning) implementation of the detection called Deep-EIR.

These features are then used to retrain the final layer of the network based on the new classes using a softmax function [12]. 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. 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.

### 3.6.3 Experimental Results

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 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 3.28 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

[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.




(a) Blurry frame

(b) Cecum

(c) Normal colon mucosa

(d) Polyp

(e) Tumor

(f) Z-line

Figure 3.28: Example for anatomic findings (classes) in the multi-class dataset. The classes are blurry fame, cecum, normal colon mucosa, polyp, tumor and Z-line.

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 3.10 shows the confusion matrix (a standard tool for evaluating multi-class classifiers showing the actual class compared to the detected class) for Multiclass-EIR. The results are a clear indication that the Multiclass-EIR approach performs well. An interesting insight

| | | *Detected class* | | | | | | Metrics | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Blurry frame | Cecum | Normal colon mucosa | Polyps | Tumor | Z-line | Precision | Recall Sensitivity | F1-score |
| *Actual class* | Blurry frame | **250** | 0 | 0 | 0 | 0 | 0 | 1.0 | 1.0 | 1.0 |
| | Cecum | 0 | **226** | 21 | 3 | 0 | 0 | 0.704 | 0.904 | 0.791 |
| | Normal colon mucosa | 0 | 85 | **165** | 0 | 0 | 0 | 0.85 | 0.66 | 0.743 |
| | Polyp | 0 | 10 | 8 | **226** | 6 | 0 | 0.953 | 0.904 | 0.928 |
| | Tumor | 0 | 0 | 0 | 8 | **242** | 0 | 0.975 | 0.968 | 0.971 |
| | Z-line | 0 | 0 | 0 | 0 | 0 | **250** | 1.0 | 1.0 | 1.0 |
| | | | | | | | Average | **0.914** | **0.906** | **0.91** |

Table 3.10: Confusion matrix and standard metrics for the six-class classification performance for Multiclass-EIR.

| | | *Detected class* | | | | | | Metrics | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Blurry frame | Cecum | Normal colon mucosa | Polyps | Tumor | Z-line | Precision | Recall Sensitivity | F1-score |
| *Actual class* | Blurry frame | **250** | 0 | 0 | 0 | 0 | 0 | 0.996 | 1.0 | 0.998 |
| | Cecum | 0 | **183** | 64 | 3 | 0 | 0 | 0.782 | 0.732 | 0.756 |
| | Normal colon mucosa | 0 | 34 | **197** | 19 | 0 | 0 | 0.641 | 0.788 | 0.707 |
| | Polyp | 1 | 17 | 45 | **183** | 4 | 0 | 0.875 | 0.732 | 0.797 |
| | Tumor | 0 | 0 | 1 | 4 | **245** | 0 | 0.983 | 0.98 | 0.981 |
| | Z-line | 0 | 0 | 0 | 0 | 0 | **250** | 1.0 | 1.0 | 1.0 |
| | | | | | | | Average | **0.879** | **0.872** | **0.876** |

Table 3.11: Confusion matrix and standard metrics for the six-classes detection performance evaluation for Deep-EIR.

is that normal colon mucosa is often miss-classified as cecum (cecum is also sometimes miss-classified as normal colon mucosa). Looking at the example images for cecum (figure 3.28(b)) and normal colon mucosa (figure 3.28(c)) reveals that this is not very surprising since it is even hard for a human observer to find clear indications for what is what. Furthermore, 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.

The performance of Deep-EIR, which is presented in table 3.11 can also be considered as good. Deep-EIR confuses the classes polyp and cecum more than Multiclass-EIR, but it is better in detecting normal colon mucosa. For detecting blurry frames and Z-lines it performs at the same level as Multiclass-EIR. Based on the confusion matrix for both approaches, we can see that for some classes Multiclass-EIR is better and for other classes Deep-EIR. In the future, a combination of both approaches might be interesting to research.

Comparing Multiclass-EIR and Deep-EIR using the standard metrics precision, recall/sensitivity and F1-score reveals that Multiclass-EIR outperforms Deep-EIR significantly with a precision of $0,914$, a recall of $0,906$ and a F1-score of $0.91$ for Multiclass-EIR compared to a precision of $0,879$, a recall of $0,872$ and a F1-score of $0.876$ for Deep-EIR. Both approaches are able to perform their multi-class classification within the defined border of $30$ FPS.

## 3.7 Summary

In this section, we presented our approach for a medical multimedia system called EIR targeted to detect diseases in the GI tract. The EIR system consists of the complete pipeline from annotation, over detection and automatic analysis to visualization. We demonstrated that all parts of the system are important by itself, and as a complete system.

The annotation subsystem as a starting point is important, because without good training data, we are not able to understand and solve the complex and often unexplored multimedia challenges of the medical field. Medical experts are always very busy and we often received the feedback that tools to provide and annotate data are very time consuming and complicated. Therefore, to make it possible for us to get data from medical institutions that we can use to test and develop applications and systems in the medical field, we had to develop several prototypes of annotation tools for medical experts. Furthermore, we performed research on these tools to find out, which ones are better usable and acceptable for the doctors [144, 145, 4, 49]. During this process, we also got access to the ASU Mayo dataset [168], which is the largest publicly available dataset of videos from GI examinations. With this dataset, we started to develop and implement an improved version of the search-based classifier.

Based on the ASU Mayo dataset, we could show that our method can achieve very good results for GI polyp detection. Since we also wanted to support medical doctors while they are performing colonoscopies, we started to develop a parallelized, real-time classification system for GI examinations [142, 143, 145]. We showed that the detection and automatic analysis subsystem can reach state-of-the-art performance for detection of polyps as use case. Moreover, we showed that the localization part is promising, but needs some improvements for the future. At the moment, EIR 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.

With EIR, we also successfully participated in the MICCAI challenge for polyp classification[5]. In the challenge, we positioned us in the middle field for the classification and detection parts, and we were the second best participant in the latency part [142]. In terms of system processing performance, we showed that using only CPUs to distribute the workloads is not enough. Therefore, we implemented, presented and evaluated an improved version of the EIR system, which uses heterogeneous architectures (GPU-acceleration) and device lending of GPUs. The quantitative results demonstrate that using distributed processing is the key to real-time performance and parallel analysis of multimedia data with different approaches [143, 145].

Furthermore, the improved EIR system reaches significant over-performance in terms of real-time video processing (300 FPS), which makes it possible to implement more extensions, for example, different feature extractors, classifiers and complex image processing algorithms to increase the number of detectable diseases by our system while keeping the real-time capability [142, 143, 145]. Moreover, we showed that the EIR system is able to perform multi-class classification (six different anatomic findings), and that the search-based Multiclass-EIR approach is able to outperform Deep-EIR, which is based on state-of-the-art deep learning (neural network) techniques. Nevertheless, it is important to point out that the used dataset is limited in size and that evaluations on larger amount of data are recommended as soon as the data is available. We also made the search-based classification part of our system open source [104],

[5] http://endovis.grand-challenge.org/

and contributed to the improvement of the open source library Lire that has been used for the implementation of our basic algorithm [97].

For the visualization subsystem, we presented two possible solutions that can be used by medical experts. Nevertheless, even if fulfilling the basic requirements, this part holds a lot of potential for future improvements [142]. Thus, in summary, EIR fulfills the requirements set in section 1.2. It is therefore a promising first step towards a medical multimedia system that can really help the medical sector in detection of some of the most lethal diseases, both for women and men.

# Chapter 4

# Conclusion

Researching and developing a complete medical multimedia system like EIR requires investigations in many different areas. The described work in this thesis touches therefore a set of diverse areas in order to learn and improve components in the entire chain of process. Initially, these areas did not seem related, but all the work is connected due to its high performance needs in terms of computational power and complexity, and also because of the search-based classification method that we researched and developed resulting in an accurate, high-precision, real-time disease detection pipeline for the GI tract (which is also depicted in figure 1.9).

## 4.1 Summary and Contributions

In this thesis, we presented our experiences with researching and developing a complete medical multimedia system for automatic analysis of the GI tract. During the journey to a complete system, we had to go beyond medical imaging for showing the potential of multimedia research beside well known scenarios like analysis of content on YouTube or Flickr [139]. Furthermore, we demonstrated that it is not appropriate to test one method or theory just with one use case, but with many different and diverse ones, to show its functionality and robustness. As a milestone and a final output of the thesis, we described our experience regarding how multimedia researchers can apply their knowledge in the medical field and published it in the ACM multimedia brave new idea track [139]. In addition to the EIR system [4, 49, 127, 126, 142, 143, 144, 145], this can be seen as an important contribution of the thesis.

The EIR system was planed and envisioned to be a complete medical multimedia system. To stay in scope of the thesis, we focused on the use case of GI disease detection using videos and images. We aimed to built a system that is generalizable, adaptable, efficient and accurate. As result, the most important outcome of this work is the EIR system, which reaches high accuracy for the polyp detection use case, is expandable with new use-cases and data types, runs in real-time and is at the moment being tested by medical experts for a patient study.

This thesis contributes in several areas of multimedia research. In a more compressed and summarized form, we contributed by researching and developing a medical multimedia system called EIR including annotation, detection and visualization tools that demonstrates the potential of multimedia research for the health care system. Further, we researched and developed an

efficient, generalizable content-based method to process multimedia data. We also contributed by researching how distributed processing can help to achieve real-time performance for medical multimedia workload processing. Moreover, we showed why the multimedia community should apply their research in medicine, and illustrated how multimedia technology and methods can be used in the medical field to improve work flows, patient care and most important potentially save lives. We also implemented and presented several prototypes and demos of the system and made parts of it open source. Finally, we contributed by writing and publishing several research papers about our findings and experiences, which we shared with the multimedia community.

All main contributions of the thesis are supported by publications in top tier conferences or journals. In more detail, the contributions in coherence to the objectives defined in section 1.2 of the thesis are:

- **Contributions to the main objective:** We developed the EIR system [142, 143, 127, 145, 126] for automatic detection of lesions in the GI tract. The system consists of an annotation, a detection and localization and a visualization subsystem. This 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.

  Using the ASU Mayo dataset [168], we showed that EIR reaches high performance in terms of both accuracy and processing. For the classification part, we can report a sensitivity of almost 98% and a precision of almost $94\%$. This means that EIR is able to find polyps in almost all cases with a high precision. This can help the medical experts to save time and lives [142, 143, 127, 145, 126]. We could also show that the EIR system is able to perform multi-class classification and that the search-based Multiclass-EIR approach is able to outperform Deep-EIR, which is based on state-of-the-art deep learning techniques. Nevertheless, it is important to point out that the used dataset is limited in size and that evaluations on larger amount of data are recommended as soon as the data is available.

  Moreover, we compared EIR with other existing 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 [142, 126, 145].

  For each part of the EIR system, we developed working prototypes and demo applications. These prototypes and demo applications have been presented at conferences [4, 142, 126, 145].

  For the real-time processing challenge, we showed that EIR 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 [142]. We researched and implemented different ways of distributed and parallel processing by using different architectures to improve the performance of the EIR system. One of the methods that we researched is the distribution of the detection and localization part on GPUs [127, 145]. Another method

that we researched was to distribute the EIR workloads via device lending [74, 126]. Both methods improved the processing performance significantly [74, 126].

We showed the potential of multimedia research in the medical field and showed possible further directions and research topics using the EIR system as an example use case [139].

We contributed to two open source projects: *Lire*, in the field of content-based image retrieval [97], and *OpenVQ*, on video quality [157]. We also released the base algorithm of EIR as an open source project called Opensea [104].

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 EIR, we researched several prototypes and techniques to make it easier and more efficient for the medical experts to transfer their knowledge to our system. For this, we explored and developed semi-supervised and cluster-based annotation tools [4, 144]. Based on the findings of one of our annotation tools, we developed a model that can be used to understand events in endoscopic surgery videos better than before and annotate this videos more efficient [49].

- **Contributions to sub-objective 2:** As the basis for the EIR system, we developed a search-based classification algorithm that uses global image features, reaches good classification performance and is very fast at the same time [136]. We developed the theory of intentional framing, which can help to explain why people take pictures and what they want to achieve with them [136]. We researched a method that can be used to accept or discard crowdsourcing workers for content annotation tasks by combining search-based classifiers with crowdsourcing information [141]. We created and researched a prototype of an intent-based video streaming system that uses the intentional framing method to save bandwidth and preserve quality of experience for video streaming [131]. We researched how the search-based classifier can be used to detect and synchronize events in image collections [196, 195]. We researched how the context (a certain watching situation) influences the quality of experience for users when they are watching videos. As a use case, we started with watching videos during a flight. We hosted a MediaEval benchmark task [138] about this topic and published a dataset [137]. Based on the use cases addressed in the thesis and the EIR system itself, we showed that the search-based classification algorithm is well suited to be applied to several different use cases that involve image classification problems [136, 141, 131, 196, 195, 138, 137, 142, 143, 127, 145, 126].

- **Contributions to sub-objective 3:** We researched different types of visualization for the output of the EIR system. The visualization includes a specific, for research and medical experts developed application [4] and an easier-to-use, web-based version [4, 145]. The visualization approaches can visualize all possible outputs of the EIR system [142].

Apart from the main contributions, we also contributed to other multimedia research relevant topics: We researched how multimedia and art can be combined to make people understand disabled people in a better way by developing a game that allows the player to experience a house from a blind person's point-of-view [140]. We developed a serious game that can simulate

the functionality of an eye-tracking device. Based on a crowdsourcing study, we could show that the data obtained by the game can be used to find areas of interest in images [134]. We also published the data obtained in this study as a publicly available dataset [133]. We researched how serious games can be used to make scientific content better accessible for the broader population. Therefore, we developed two game prototypes and tested them with real users [98, 108]. We explored how multimedia methods can be used to find manipulations in online content like images and videos, and how to verify that the content has not been manipulated [13]. We also published a dataset that we collected during this study [26]. We researched how the design of complex crowdsourcing tasks can influence their outcome on the use case of 3D reconstruction of soccer players. We published a best practice paper about design and test of such tasks [191] and published a dataset obtained in the study [132]. We looked into the problem of how crowdsourcing can be used in subjective studies such as quality of experience in videos. For this, we looked at different tiling strategies in a football video streaming system [42]. Based on this research, we also investigated how too much control can have a negative influence on a crowdsourcing study and reported our findings, showing that crowdsourcing can lead to a self-fulfilling prophecy [135].

In addition to the above contributions, the author also supervised several master students, organized workshops and was part of program committees for conferences. We also collaborated with the Cancer registry of Norway in a project that tries to increase awareness for HPV and cervical cancer. The cancer registry of Norway started a big user study based on the application, which we helped to develop in December 2015 [108, 98].

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, researchers can improve the EIR system in several ways and extend it with new technologies and methods like more sophisticated deep learning approaches, pre-processing of images and videos and including more sources of data such as patient records. Other improvements can be a more detailed comparision of standard machine learning methods and deep learning and when it would be reasonable to switch between them or if even a combination of both is possible. Furthermore, we did not look into the time dimension in the videos, which is another source of information that could help to improve the accuracy of the automatic analysis.

Another important point is to collect more training data in the medical field. A large, but not yet annotated, dataset that has been collected during this work, holds a lot of challenges and 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, it is not completed yet. This will also enable researchers to further investigate and evaluate the

multi-class classification part in more detail. Another interesting direction for future work is the localization part of the system. The researched and developed method is able to localize polyps, but it would also be interesting to extend it to other diseases. Apart from the extension to more diseases, it would also be interesting to investigate the potential of deep learning for the localization of endoscopic findings.

Further improvements of the automatic analysis and localisation could most probably be achieved with performing 3D reconstruction of the GI tract. A 3D representation of the GI tract could make it easier to detect diseases, and it would also enable size estimation of, for example, polyps, which is an important information for doctors.

The output of an automatic system like EIR also opens many possibilities for visualization, automatic reporting and computer aided diagnosis application scenarios. For example, one could use the automatic output of EIR to add information to patient records such as images of the found diseases or video clips. Moreover, the automatic analysis can also be used to create automatic reports after the examination that could help medical doctors to reduce the amount of time spend on reporting. The saved time could then be used to perform additional examinations.

Finally, a clinical trial in collaboration with medical doctors would also be an interesting and challenging topic. We showed with our work about cervical cancer [98, 108] and the brave new idea paper [139] that the possibilities for multimedia research in the field of medicine are endless and other medical fields like psychology or pregnancy ultrasound, etc., can also be interesting directions for researchers.

## 4.3 Final Remarks

To make it possible to continue the research in medical multimedia systems, we have applied for several projects at the Norwegian research council. Two of them got funded and gave us a lot of working hours from medical experts to create better datasets. The future plan is to make this medical multimedia data and medical expertise publicly available. Furthermore, we applied together with the Drammen hospital and the Cancer Registry of Norway, for an innovation project at Helse Sør-Øst to build a live system for colonoscopies. This system will be based on our current EIR system and has many system research challenges to tackle, i.e., it has to work in real-time, preserve privacy and be fault tolerant. We fulfilled all research goals that we specified for this thesis and created a complete system that can be used as basis for future research and most important has the potential to 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 1.9.

## 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:** Riegler contributed to the text in all sections and performed some performance measurements for the speed of the system. He measured and reported the performance of Lire regarding the search based algorithm. Furthermore, he contributed in the development of the demo application and described a medical use case for it in the field of gastrointestinal analysis.

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

**Contributed to:** Main Objective, Sub-objective 2

## 5.2 Paper II: How 'How Reflects What's What: Content-based Exploitation of How Users Frame Social Images

**Authors:** Michael Riegler, Martha Larson, Mathias Lux, Christoph Kofler

**Abstract:** In this paper, we introduce the concept of intentional framing, defined as the sum of the choices that a photographer makes on how to portray the subject matter of an image. We carry out analysis experiments that demonstrate the existence of a correspondence between image similarity that is calculated automatically on the basis of global feature representations, and image similarity that is perceived by humans at the level of intentional frames. Intentional framing has profound implications: The existence of a fundamental image-interpretation principle that explains the importance of global representations in capturing human perceived image semantics reaches beyond currently dominant assumptions in multimedia research. The ability of fast global-feature approaches to compete with more 'sophisticated' approaches, which are computationally more complex, is demonstrated using a simple search method (Sim- Sea) to classify a large (2M) collection of social images by tag class. In short, intentional framing provides a principled connection between human interpretations of images and lightweight, fast image processing methods. Moving forward, it is critical that the community explicitly exploits such approaches, as the social image collections that we tackle, continue to grow larger.

**Author's contributions:** Riegler had the overall responsibility and the idea for the paper. Riegler wrote most of the text and performed all experiments. He developed the idea of intentional framing together with Martha Larson an Mathias Lux. He implemented the search based classification algorithm using the open source software Lire as platform. He tested it on three different use cases, conducted experiments and analyzed and reported the results. He also performed extensive literature research to support the theory. He performed tests that compare the performance of global features with local features.

**Published in:** ACM Multimedia Conference (MM), 2014.

**Contributed to:** Main Objective, Sub-objective 2

## 5.3 Paper III: Exploitation of Producer Intent in Relation to Bandwidth and QoE for Online Video Streaming Services

**Authors:** Michael Riegler, Lilian Calvet, Amandine Calvet, Pål Halvorsen, Carsten Griwodz

**Abstract:** This paper is the product of recent advances in research on users' intent during multimedia content retrieval. Our goal is to save bandwidth while streaming video clips from a browsable on-demand service, while maintaining or even improving the users' quality of experience (QoE). Understanding user intent allows us to predict whether streaming a particular video in a low quality constitutes a reduced QoE for a user. However, many

VoD streaming services today are used by users for a wide variety of reasons, meaning that user intent cannot be inferred from their use of the service alone. However, our investigation demonstrates that user intent does in most cases coincide with producer intent. We can also demonstrate that the latter can be inferred from the content itself as well as associated metadata. By transitivity, we can choose a default video quality that satisfies the users QoE in the majority of cases.

**Author's contributions:** Riegler brought the original idea and concept for the paper. Riegler had the overall responsibility for the writing process. He designed and implemented the proposed system. The system is based on the intentional framing idea and uses different types of information to classify videos into different intent classes. He also conducted the user study. For the user study he developed a web based questionnaire. He collected the user feedback and analyzed the results. He also presented a analysis that shows how much bandwidth can be saved using the proposed system.

**Published in:** ACM SIGMM Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV), 2015.

**Contributed to:** Main Objective, Sub-objective 2

## 5.4 Paper IV: Media Synchronization and Sub-Event Detection in Multi-User Image Collections

**Authors:** Maia Zaharieva, Michael Riegler

**Abstract:** Personal media capturing devices, such as smartphones or personal image and video cameras, are rarely synchronized. As a result, common tasks, like event detection and summarization across different multi-user media galleries, are considerably impeded and error-prone. In this paper, we investigate different approaches for the synchronization of image collections using visual information only. We perform a thorough evaluation of the performance of several global features on three datasets. Additionally, we explore the feasibility of common clustering algorithms for the detection of sub-events in the presence of synchronization misalignment.

**Author's contributions:** Riegler contributed to the text and with the results and analysis of the search based classification algorithm applied to the event synchronization problem. He also contributed to the idea and concept of the paper. He modified the search based algorithm so that it can be used for event detection by adding a functionality that can be used to cluster different events in same clusters. He performed the experiments for the x-means and x-means & time approaches. He helped analyzing the results of the experiment and draw conclusions. He contributed with text in all sections.

**Published in:** ACM Human Event Workshop at ACM Multimedia Conference (HuEvent), 2015.

**Contributed to:** Main Objective, Sub-objective 2

## 5.5 Paper V: Multimodal Synchronization of Image Galleries

**Authors:** Maia Zaharieva, Michael Riegler, Manfred Del Fabro

**Abstract:** This paper describes our contribution to the MediaEval 2014 task on the Synchronization of multi-user EventMedia (SEM). We propose two multimodal approaches that employ both visual and time information for the synchronization of different images galleries and for the detections of sub-events. The methods prove robustness in the determination of time offsets with accuracy of up to 87%.

**Author's contributions:** Riegler contributed by running experiments using global image features and a search based method based clustering. He run experiments for run 3 and run 4 presented in the paper. He described his methods and contributed to the analysis and conclusion. He contributed to the writing and developing the idea.

**Published in:** MediaEval Workshop, 2014.

**Contributed to:** Main Objective, Sub-objective 2

## 5.6 Paper VI: Introduction to a Task on Context of Experience: Recommending Videos Suiting a Watching Situation

**Authors:** Michael Riegler, Martha Larson, Concetto Spampinato, Jonas Markussen, Pål Halvorsen, Carsten Griwodz

**Abstract:** We propose a Context of Experience task, whose aim it is to explore the suitability of video content for watching in certain situations. Specifically, we look at the situation of watching movies on an airplane. As a viewing context, airplanes are characterized by small screens and distracting viewing conditions. We assume that movies have properties that make them more or less suitable to this context. We are interested in developing systems that are able to reproduce a general judgment of viewers about whether a given movie is a good movie to watch during a flight. We provide a data set including a list of movies and human judgments concerning their suitability for airplanes. The goal of the task is to use movie metadata and audio-visual features extracted from movie trailers in order to automatically reproduce these judgments. A basic classification system demonstrates the feasibility and viability of the task.

**Author's contributions:** Riegler brought the original idea and concept for the paper. Together with Larson he refined the idea and decided for a use case (movies on flights). He collected the data by crawling airline homepages to get a subset of movies. Together with Spampinato he decided with additional data should be collected for each movie. He performed initial experiments using the Weka library and different subsets of data. He presented and discussed the results in the paper. He wrote most parts of the text.

**Published in:** MediaEval Workshop, 2015.

**Contributed to:** Main Objective, Sub-objective 2

## 5.7 Paper VII: Right inflight? A Dataset for Exploring the Automatic Prediction of Movies Suitable for a Watching Situation

**Authors:** Michael Riegler, Martha Larson, Concetto Spampinato, Pål Halvorsen, Mathias Lux, Jonas Markussen, Konstantin Pogorelov, Carsten Griwodz, Håkon Stensland

**Abstract:** In this paper, we present the dataset Right Inflight developed to support the exploration of the match between video content and the situation in which that content is watched. Specifically, we look at videos that are suitable to be watched on an airplane, where the main assumption is that that viewers watch movies with the intent of relaxing themselves and letting time pass quickly, despite the inconvenience and discomfort of flight. The aim of the dataset is to support the development of recommender systems, as well as computer vision and multimedia retrieval algorithms capable of automatically predicting which videos are suitable for inflight consumption. Our ultimate goal is to promote a deeper understanding of how people experience video content, and of how technology can support people in finding or selecting video content that supports them in regulating their internal states in certain situations. Right Inflight consists of 318 human-annotated movies, for which we provide links to trailers, a set of pre-computed low-level visual, audio and text features as well as user ratings. The annotation was performed by crowdsourcing workers, who were asked to judge the appropriateness of movies for inflight consumption.

**Author's contributions:** Riegler brought the original idea and concept for the paper. He collected more data by conducting a crowdsourcing study asking people for their opinion about movies. He performed experiments using the Weka library and different subsets of data, specifically focused on global image features compared to metadata approaches. He presented and discussed the results in the paper. He wrote most parts of the text. He cleaned the data and made it public available for the dataset paper.

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

**Contributed to:** Main Objective, Sub-objective 2

## 5.8 Paper VIII: Expert Driven Semi-Supervised Elucidation Tool for Medical Endoscopic Videos

**Authors:** Zeno Albisser, Michael Riegler, Pål Halvorsen, Jiang Zhou, Carsten Griwodz, Ilangko Balasingham, Cathal Gurrin

**Abstract:** In this paper, we present a novel application for elucidating all kind of videos that require expert knowledge, e.g., sport videos, medical videos etc., focusing on endoscopic

surgery and video capsule endoscopy. In the medical domain, the knowledge of experts for tagging and interpretation of videos is of high value. As a result of the stressful working environment of medical doctors, they often simply do not have time for extensive annotations. We therefore present a semisupervised method to gather the annotations in a very easy and time saving way for the experts and we show how this information can be used later on.

**Author's contributions:** Riegler brought the original idea and the main concept for the paper. which was based on another tool that has been developed from Riegler before. Riegler provided the source code for the HTML5 version of the tool based on the previous application developed by him. He contributed also in the writing of the paper and with discussions and advices. He also contributed with his knowledge in the medical field of endoscopic surgeries.

**Published in:** ACM Multimedia Systems Conference (MMSys), 2015.

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

## 5.9 Paper IX: Event Understanding in Endoscopic Surgery Videos

**Authors:** Mario Guggenberger, Michael Riegler, Mathias Lux, Pål Halvorsen

**Abstract:** Event detection and understanding is an important area in computer science and especially multimedia. The term event is very broad, and we want to propose a novel event based view on endoscopic surgeries. Thus, with the novel view on surgery in this paper, we want to provide a better understanding and possible way of segmentation of the whole event surgery but also the included sub-events. To achieve this sophisticated goal, we present an annotation tool in combination with a thinking aloud test with an experienced surgeon.

**Author's contributions:** Riegler contributed in the design of the user study and did also the main parts of the writing. He did the conceptual analysis of the interview results by reviewing the video material and the questionnaire. Based on this he developed the model of describing events in endoscopic surgeries. Together with Guggenberger he combined the technical and medical findings to draw a final conclusion.

**Published in:** ACM Human Event Workshop at ACM Multimedia Conference (HuEvent), 2015.

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

## 5.10 Paper X: 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:** Riegler had the main idea for the paper and was responsible for the writing and the coordination. He performed the experiments to evaluate the performance of the clustering application. Therefore, he programmed a clustering application and conducted tests on the intent dataset and the ASU Mayo dataset for polyp classification. He also discussed the results of these experiments in the paper and concluded them. The application is based on the x-means clustering. The basic code of this application has been used by Pogorelov to develop the tree based representation of the clustering output and the annotation part of it.

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

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

## 5.11 Paper XI: 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:** This is the first paper describing the EIR system. The main concept of the EIR system has been developed by Riegler during his PhD. He was mainly responsible for coordinating the writing process and the experiments. The also worked closely with the medical experts that contributed to the paper to create a solid medical basic for the use case. He designed and developed the main architecture of the system. He developed and conducted the experiments for the detection part of the system. He performed extensive research of the related work. He described the basic idea and the real world scenarios. He also performed tests in terms of speed of the system.

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

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

## 5.12 Paper XII: 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:** This paper is an extension of the basic EIR paper. Riegler was responsible for planning it and the writing process. He also addressed most of the remarks of the reviewers after the first revision. He extended all sections in the paper either with

new information and text or additional experiments. He extended the basic idea description of the system for overall description, the annotation, visualization and the detection part. He extended the experimental part with detailed accuracy experiments for the detection part. He also extended the experiments with a detailed performance evaluation in terms of speed and memory consumption for the detection part. Riegler also presented together with Pogorelov the results for the MICCAI challenge on polyp detection where a combination of the localization and the detection part has been used to participate. He extended the related work section with more details and added an additional section about neural networks. Finally, he contributed by extending the real world use case section and discussion of the experimental results.

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

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

## 5.13 Paper XIII: Computer Aided Disease Detection System for Gastrointestinal Examinations

**Authors:** Michael Riegler, Konstantin Pogorelov, Jonas Markussen, Mathias Lux, Håkon Kvale Stensland, Thomas de Lange, Carsten Griwodz, Pål Halvorsen, Dag Johansen, Peter T. Schmidt, Sigrun L. Eskeland

**Abstract:** In this paper, we present the computer-aided diagnosis part of the EIR system, which can support medical experts in the task of detecting diseases and anatomical landmarks in the gastrointestinal (GI) system. This includes automatic detection of important findings in colonoscopy videos and marking them for the doctors. EIR is designed in a modular way so that it can easily be extended for other diseases. For this demonstration, we will focus on polyp detection, as our system is trained with the ASU Mayo Clinic polyp database.

**Author's contributions:** Riegler was the main responsible person for the paper. He coordinated the writing process and designed the concept of the paper. He mainly contributed to the introduction, related work and basic description of the detection subsystem. He contributed also with an experiment based on the ASU Mayo data that showed the classification performance of the system.

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

**Contributed to:** Main Objective, Sub-objective 2

## 5.14 Paper XIV: 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:** Riegler was the coordinator of the paper and also mainly responsible for the writing. The paper got first a conditional accept and had to be rewritten to fulfil the brave new idea requirements of the conference. Riegler contributed by conducting an extensive literature research for the related work. He also performed preliminary experiments and presented them in the results. Furthermore, the contributed by pointing out multimedia challenges in the health care field by using the EIR system as an example. He also discussed related work in context to new trends. He also contributed most of the outlook and challenges discussion. Finally, he contributed to the open challenges discussion. We also have to mention that we had a shepherd during the writing process (Martha Larson) that helped to improve the quality of the paper and the focus so that it fits the brave new idea track.

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

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

## 5.15 Paper XV: 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:** Riegler had the conceptual idea and the lead of the paper. He was responsible for the writing process. He contributed mainly by writing text and analysis of the experiments conducted by Pogorelov. He also provided a Java implementation of the detection subsystem that has been improved by Pogorelov with adding GPU support.

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

**Contributed to:** Main Objective, Sub-objective 2

## 5.16 Paper XVI: Device Lending in PCI Express Networks

**Authors:** Lars Bjorlykke Kristiansen, Jonas Markussen, Håkon Kvale Stensland, Michael Riegler, Hugo Kohmann, Friedrich Seifert, Roy Nordstrøm, Carsten Griwodz, Pål Halvorsen

**Abstract:** The challenge of scaling IO performance of multimedia systems to demands of their users has attracted much research. A lot of effort has gone into development of distributed systems that add little latency and computing overhead. For machines in PCI Express (PCIe) clusters, we propose Device Lending as a novel solution which works at a system level. Device Lending achieves low latency and extremely low computing overhead without requiring any application-specific distribution mechanisms. For applications, the remote IO resource appears local. In fact, even the drivers of the operating system remain unaware that hardware resources are located in remote machines. By enabling machines in a PCIe cluster to lend a wide variety of hardware, cluster machines can

get temporary access to a pool of IO resources. Network cards, FPGAs, SSDs, and even GPUs can easily be shared among computers. Our pro- posed solution, Device Lending, works transparently with- out requiring any modifications to drivers, operating systems or software applications.

**Author's contributions:** Riegler contributed mainly with some input in the related work and by reviewing especially the abstract and the introduction. He also reviewed the paper and gave feedback. Riegler also had several discussions with Markussen about the paper. The main contribution was to lead a paper for a demo that shows how device lending can be used for multimedia workloads. The paper is referenced as example in this paper.

**Published in:** ACM SIGMM Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV), 2016.

**Contributed to:** Main Objective, Sub-objective 2

## 5.17 Paper XVII: 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:** Riegler developed the idea for the paper together with Markussen and Pogorelov. He had the lead for the writing process and the collaboration between the authors. He contributed the introduction, the use case scenario, the abstract and the conclusion. He contributed by going to the hospital and taking pictures and talking to doctors to be able to describe the use case better. He also helped to plan and analyse the experiment conducted by Pogorelov and Markussen. He was reviewing the paper several times and made improvements in the text.

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

**Contributed to:** Main Objective, Sub-objective 2

# Bibliography


[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015. *Software available from tensorflow. org*, 1, 2015.

[2] D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.

[3] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 510–517. IEEE, 2012.

[4] 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.

[5] 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.

[6] G. Amato and F. Falchi. knn based image classification relying on local feature similarity. In *Proceedings of the Third International Conference on Similarity Search and Applications*, pages 101–108. ACM, 2010.

[7] 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.

[8] L. Arge. The buffer tree: A new technique for optimal i/o-algorithms. In *Workshop on Algorithms and Data structures*, pages 334–345. Springer, 1995.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[10] J. Bernal, J. Sánchez, and F. Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

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

[12] C. M. Bishop. Pattern recognition. *Machine Learning*, 128, 2006.

[13] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, M. Larson, and Y. Kompatsiaris. Verifying multimedia use at mediaeval 2015. In *Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany, 09/2015 2015. CEUR Workshop Proceedings.

[14] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via plsa. In *European conference on computer vision*, pages 517–530. Springer, 2006.

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

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

[17] D. T. Bui, T. A. Tuan, H. Klempe, B. Pradhan, and I. Revhaug. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, 13(2):361–378, 2016.

[18] 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):1551–1564, 2012.

[19] 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.

[20] S. Chatzichristofis, Y. Boutalis, and M. Lux. Selection of the proper compact composite descriptor for improving content based image retrieval. In *Proceedings of the 6th IASTED International Conference*, volume 134643, page 64, 2009.

[21] 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.

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

[23] 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.

[24] C. Chin and D. E. Brown. Learning in science: A comparison of deep and surface approaches. *Journal of research in science teaching*, 37(2):109–138, 2000.

[25] V. Conotter, D.-T. Dang-Nguyen, G. Boato, M. Menéndez, and M. Larson. Assessing the impact of image manipulation on users' perceptions of deception. In *IS&T/SPIE Electronic Imaging*, pages 90140Y–90140Y. International Society for Optics and Photonics, 2014.

[26] V. Conotter, D.-T. Dang-Nguyen, M. Riegler, G. Boato, and M. Larson. A crowdsourced data set of edited images online. In *CrowdMM '14: Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, page 49–52, New York, NY, USA, 2014. ACM.

[27] V. Dasigi, R. C. Mann, and V. A. Protopopescu. Information fusion for text classification: An experimental comparison. *Pattern Recognition*, 34(12):2413–2425, 2001.

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

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[30] 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.

[31] 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 *ICML*, pages 647–655, 2014.

[32] G. H. Dunteman. *Principal components analysis*, volume 69. Sage, 1989.

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

[34] H. J. Escalante, C. A. Hérnadez, L. E. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 172–179. ACM, 2008.

[35] H. Espeland, P. B. Beskow, H. K. Stensland, P. N. Olsen, S. Kristoffersen, C. Griwodz, and P. Halvorsen. P2g: A framework for distributed real-time processing of multimedia data. In *Proceedings of the 2011 40th International Conference on Parallel Processing Workshops*, pages 416–426. IEEE, 2011.

[36] H. Fang, Z. Zhang, C. J. Wang, M. Daneshmand, C. Wang, and H. Wang. A survey of big data research. *IEEE network*, 29(5):6, 2015.

[37] H. G. Feichtinger and T. Strohmer. *Gabor analysis and algorithms: Theory and applications*. Springer, 1998.

[38] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. W. Coebergh, H. Comber, D. Forman, and F. Bray. Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012. *Eur J Cancer*, 49(6):1374–403, 2012.

[39] Flickr. Flickr largest image size. http://www.flickr.com/help/photos/. [last visited, Oct., 2016].

[40] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[41] M. Frigo, C. E. Leiserson, H. Prokop, and S. Ramachandran. Cache-oblivious algorithms. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 285–297. IEEE, 1999.

[42] V. R. Gaddam, M. Riegler, R. Eg, C. Griwodz, and P. Halvorsen. Tiling in interactive panoramic video: Approaches and evaluation. *IEEE Transactions on Multimedia*, 18(9):1819–1831, 2016.

[43] I. Gialampoukidis, A. Moumtzidou, D. Liparas, S. Vrochidis, and I. Kompatsiaris. A hybrid graph-based and non-linear late fusion approach for multimedia retrieval. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2016.

[44] 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.

[45] B. Goetz and T. Peierls. *Java concurrency in practice*. Pearson Education, 2006.

[46] P. H. Gosselin and M. Cord. A comparison of active classification methods for content-based image retrieval. In *Proceedings of the 1st international workshop on Computer vision meets databases*, pages 51–58. ACM, 2004.

[47] H. Greenspan, B. van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

[48] A. Guérin-Dugué and A. Oliva. Classification of scene photographs from local orientations features. *Pattern Recognition Letters*, 21(13):1135–1140, 2000.

[49] M. Guggenberger, M. Riegler, M. Lux, and P. Halvorsen. Event understanding in endoscopic surgery videos. In *Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia*, pages 17–22. ACM, 2014.

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

[51] A. H. Gunatilaka and B. A. Baertlein. Feature-level and decision-level fusion of non-coincidently sampled sensors for land mine detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):577–589, 2001.

[52] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[53] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 13–16. IEEE, 1992.

[54] 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.

[55] P. Halvorsen, S. Sægrov, A. Mortensen, D. K. Kristensen, A. Eichhorn, M. Stenhaug, S. Dahl, H. K. Stensland, V. R. Gaddam, C. Griwodz, et al. Bagadus: an integrated system for arena sports analytics: a soccer case study. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 48–59. ACM, 2013.

[56] J. Han and K.-K. Ma. Fuzzy color histogram and its use in color image retrieval. *IEEE Transactions on image Processing*, 11(8):944–952, 2002.

[57] R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[58] T. Hastie, R. Tibshirani, and J. Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

[59] D. M. Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[60] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[61] J. Hays and A. A. Efros. Scene completion using millions of photographs. *Communications of the ACM*, 51(10):87–94, 2008.

[62] C. Hewitt. Viewing control structures as patterns of passing messages. *Artificial intelligence*, 8(3):323–364, 1977.

[63] W. D. Hillis and G. L. Steele Jr. Data parallel algorithms. *Communications of the ACM*, 29(12):1170–1183, 1986.

[64] 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. *Cochrane Database Syst Rev*, 9:Cd009259, 2013.

[65] P. Howarth and S. Rüger. Evaluation of texture features for content-based image retrieval. In *Image and Video Retrieval*, pages 326–334. Springer, 2004.

[66] D. F. Hsu and I. Taksa. Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval*, 8(3):449–480, 2005.

[67] C. J. Hughes, P. Kaul, S. V. Adve, R. Jain, C. Park, and J. Srinivasan. Variability in the execution of multimedia applications and implications for architecture. In *Proceedings of the Computer Architecture, 2001 28th Annual International Symposium on*, pages 254–265. IEEE, 2001.

[68] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. De Groen. Polyp detection in colonoscopy video using elliptical shape feature. In *2007 IEEE International Conference on Image Processing*, volume 2, pages II–465. IEEE, 2007.

[69] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[70] 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.

[71] J. Kang and R. Doraiswami. Real-time image processing system for endoscopic applications. In *Electrical and Computer Engineering, 2003. IEEE CCECE 2003. Canadian Conference on*, volume 3, pages 1469–1472. IEEE, 2003.

[72] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[73] D. E. Knuth. *The art of computer programming: sorting and searching*, volume 3. Pearson Education, 1998.

[74] 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.

[75] J. Krüger and R. Westermann. Linear algebra operators for gpu implementation of numerical algorithms. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 908–916. ACM, 2003.

[76] H. Kuchen and M. Cole. The integration of task and data parallel skeletons. *Parallel Processing Letters*, 12(02):141–155, 2002.

[77] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. Jones. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the 1st ACM international conference on multimedia retrieval*, page 51. ACM, 2011.

[78] C. Lattner and V. Adve. Llvm: A compilation framework for lifelong program analysis & transformation. In *Code Generation and Optimization, 2004. CGO 2004. International Symposium on*, pages 75–86. IEEE, 2004.

[79] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[80] T.-W. Lee. *Independent component analysis*. Springer, 1998.

[81] V. W. Lee, C. Kim, J. Chhugani, M. Deisher, D. Kim, A. D. Nguyen, N. Satish, M. Smelyanskiy, S. Chennupaty, P. Hammarlund, et al. Debunking the 100x gpu vs. cpu myth: an evaluation of throughput computing on cpu and gpu. *ACM SIGARCH Computer Architecture News*, 38(3):451–460, 2010.

[82] D. Leijen, W. Schulte, and S. Burckhardt. The design of a task parallel library. *ACM Sigplan Notices*, 44(10):227–242, 2009.

[83] S. Leutenegger, M. Chli, and R. Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE, 2011.

[84] 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.

[85] 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.

[86] X. Li, L. Chen, L. Zhang, F. Lin, and W.-Y. Ma. Image annotation by large-scale content-based image retrieval. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 607–610. ACM, 2006.

[87] X. Li, C. G. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 10–17. ACM, 2010.

[88] Y. Li, D. J. Crandall, and D. P. Huttenlocher. Landmark classification in large-scale image collections. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1957–1964. IEEE, 2009.

[89] S. Liang. *The Java Native Interface: Programmer's Guide and Specification*. Addison-Wesley Professional, 1999.

[90] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: fast feature extraction and svm training. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1689–1696. IEEE, 2011.

[91] T. A. Longacre and C. M. Fenoglio-Preiser. Mixed hyperplastic adenomatous polyps/serrated adenomas: a distinct form of colorectal neoplasia. *The American journal of surgical pathology*, 14(6):524–537, 1990.

[92] B. Loni, J. Hare, M. Georgescu, M. Riegler, X. Zhu, M. Morchid, R. Dufour, and M. Larson. Getting by with a little help from the crowd: Practical approaches to social image labeling. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, pages 69–74. ACM, 2014.

[93] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[94] Z. Lu, J. Hein, M. Humphrey, M. Stan, J. Lach, and K. Skadron. Control-theoretic dynamic frequency and voltage scaling for multimedia workloads. In *Proceedings of the 2002 international conference on Compilers, architecture, and synthesis for embedded systems*, pages 156–163. ACM, 2002.

[95] M. Lux. Lire: open source image retrieval in java. In *Proceedings of the 21st ACM MM*, pages 843–846. ACM, 2013.

[96] M. Lux and O. Marques. Visual information retrieval using java and lire. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 5(1):1–112, 2013.

[97] 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.

[98] M. Machnik, M. Riegler, and S. Sen. Crowdpinion: Motivating people to share their momentary opinion. In *Proceedings of the GamifIR Workshop at ECIR*, pages 44–51, 2015.

[99] S. Mallery and J. Van Dam. Advances in diagnostic and therapeutic endoscopy. *Med Clin North Am*, 84(5):1059–83, 2000.

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

[101] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers. Big data full report. `http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx`. [last visited, Oct. 12, 2016].

[102] Marie Moen Kingsrod. Tarmkreft: Den tause folkesykdommen. `http://pluss.vg.no/2016/02/03/2300/2300_23607320`. [last visited, Oct. 12, 2016].

[103] K. Mc Donald and A. F. Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *Image and video retrieval*, pages 61–70. Springer, 2005.

[104] Media Performance Group. Opensource Project: Opensea - Search-based classification. `https://bitbucket.org/mpg_projects/opensea`. [last visited, Oct. 12, 2016].

[105] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[106] T. M. Mitchell. Machine learning. *Artificial Intelligence*, 1997.

[107] S. Mittal and J. S. Vetter. A survey of cpu-gpu heterogeneous computing techniques. *ACM Computing Surveys (CSUR)*, 47(4):69, 2015.

[108] W. Moazzam, M. Riegler, S. Sen, and M. Nygård. Scientific hangman: Gamifying scientific evidence for general public. In *Proceedings of the GamifIR Workshop at ECIR*, pages 26–33, 2015.

[109] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[110] B. Münzer, K. Schoeffmann, and L. Böszörmenyi. Detection of circular content area in endoscopic videos. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 534–536. IEEE, 2013.

[111] B. Münzer, K. Schoeffmann, and L. Böszörmenyi. Improving encoding efficiency of endoscopic videos by using circle detection based border overlays. In *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, pages 1–4. IEEE, 2013.

[112] B. Münzer, K. Schoeffmann, and L. Böszörmenyi. Relevance segmentation of laparoscopic videos. In *Multimedia (ISM), 2013 IEEE International Symposium on*, pages 84–91. IEEE, 2013.

[113] 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.

[114] J. Norum and J. Olsen. A cost-effectiveness approach to the norwegian follow-up programme in colorectal cancer. *Annals of oncology*, 8(11):1081–1087, 1997.

[115] Nvidia, CUDA. Compute unified device architecture programming guide. 2007.

[116] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[117] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[118] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[119] P. N. Olsen, M. Nyhus, P. Halvorsen, and C. Griwodz. A logical memory model for scaling parallel multimedia workloads. In *Proceedings of the 25th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 49–54. ACM, 2015.

[120] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips. Gpu computing. *Proceedings of the IEEE*, 96(5):879–899, 2008.

[121] J. D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krüger, A. E. Lefohn, and T. J. Purcell. A survey of general-purpose computation on graphics hardware. In *Computer graphics forum*, volume 26, pages 80–113. Wiley Online Library, 2007.

[122] N. Park, B. Hong, and V. K. Prasanna. Tiling, block data layout, and memory hierarchy performance. *IEEE Transactions on Parallel and Distributed Systems*, 14(7):640–654, 2003.

[123] G. Pass and R. Zabih. Comparing images using joint histograms. *Multimedia systems*, 7(3):234–240, 1999.

[124] D. Pelleg, A. W. Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, volume 1, 2000.

[125] C. Poellabauer and K. Schwan. Kernel support for the event-based cooperation of distributed resource managers. In *Proceedings of the Real-Time and Embedded Technology and Applications Symposium, 2002 IEEE*, pages 3–12. IEEE, 2002.

[126] 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.

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

[128] S. Popat, R. Hubner, and R. Houlston. Systematic review of microsatellite instability and colorectal cancer prognosis. *Journal of Clinical Oncology*, 23(3):609–618, 2005.

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

[130] P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. In *Encyclopedia of database systems*, pages 532–538. Springer, 2009.

[131] M. Riegler, L. Calvet, A. Calvet, P. Halvorsen, and C. Griwodz. Exploitation of producer intent in relation to bandwidth and qoe for online video streaming services. In *Proceedings of the 25th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 7–12. ACM, 2015.

[132] 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.

[133] M. Riegler, R. Eg, L. Calvet, M. Lux, P. Halvorsen, and C. Griwodz. Playing around the eye tracker-a serious game based dataset. In *Proceedings of the GamifIR Workshop at ECIR*, pages 34–40, 2015.

[134] M. Riegler, R. Eg, M. Lux, and M. Schicho. Mobile picture guess: A crowdsourced serious game for simulating human perception. In *International Conference on Social Informatics*, pages 461–468. Springer, 2014.

[135] M. Riegler, V. R. Gaddam, M. Larson, R. Eg, P. Halvorsen, and C. Griwodz. Crowdsourcing as self-fulfilling prophecy: Influence of discarding workers in subjective assessment tasks. In *Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on*, pages 1–6. IEEE, 2016.

[136] 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*, pages 397–406. ACM, 2014.

[137] 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.

[138] M. Riegler, M. Larson, C. Spampinato, J. Markussen, P. Halvorsen, and C. Griwodz. Introduction to a task on context of experience: Recommending videos suiting a watching situation. In *Proceedings of the MediaEval 2015 Workshop*. CEUR Workshop Proceedings, 2015.

[139] M. Riegler, M. Lux, C. Gridwodz, 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.

[140] M. Riegler, M. Lux, C. Zellot, L. Knoch, H. Schnattler, S. Napetschnig, J. Kogler, C. Degendorfer, and M. Zoderer. Gone: an interactive experience for two people. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 701–704. ACM, 2014.

[141] M. Riegler, P. N. Olsen, and P. Halvorsen. Work like a bee-taking advantage of diligent crowdsourcing workers. In *Proceedings of the MediaEval 2014 Workshop*. CEUR Workshop Proceedings, 2014.

[142] M. Riegler, K. Pogorelov, S. L. Eskeland, P. T. Schmidt, Z. Albisser, D. Johansen, C. Griwodz, P. l. Halvorsen, and T. de Lange. From annotation to computer aided diagnosis: Detailed evaluation of a medical multimedia system. *Submitted to: ACM Transactions on Multimedia Computing, Communications and Applications*, 2016.

[143] 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.

[144] 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.

[145] 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.

[146] S. Ryoo, C. I. Rodrigues, S. S. Baghsorkhi, S. S. Stone, D. B. Kirk, and W.-m. W. Hwu. Optimization principles and application performance evaluation of a multithreaded gpu using cuda. In *Proceedings of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming*, pages 73–82. ACM, 2008.

[147] M. A. Savelonas, I. Pratikakis, and K. Sfikas. Fisher encoding of differential fast point feature histograms for partial 3d object retrieval. *Pattern Recognition*, 55:114–124, 2016.

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

[149] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[150] T. Semertzidis, D. Rafailidis, E. Tiakas, M. G. Strintzis, and P. Daras. Multimedia indexing, search, and retrieval in large databases of social networks. In *Social Media Retrieval*, pages 43–63. Springer, 2013.

[151] S. Sen, S. Chatterjee, and N. Dumir. Towards a theory of cache-efficient algorithms. *Journal of the ACM (JACM)*, 49(6):828–858, 2002.

[152] H. Shachnai and S. Y. Philip. *The role of wait tolerance in effective batching: A paradigm for multimedia scheduling schemes*. IBM TJ Watson Research Center, 1995.

[153] 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.

[154] T. Sikora. The mpeg-7 visual standard for content description-an overview. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):696–702, 2001.

[155] E. Sintorn and U. Assarsson. Fast parallel gpu-sorting using a hybrid algorithm. *Journal of Parallel and Distributed Computing*, 68(10):1381–1388, 2008.

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

[157] 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.

[158] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM, 2005.

[159] J. Song. Effective hashing for large-scale multimedia search. In *Proceedings of the 2013 Sigmod/PODS Ph. D. symposium on PhD symposium*, pages 55–60. ACM, 2013.

[160] A. Sonnenberg and R. M. Genta. Low prevalence of colon polyps in chronic inflammatory conditions of the colon. *The American journal of gastroenterology*, 110(7):1056–1061, 2015.

[161] Spark, Apache. Lightning-fast cluster computing, 2013.

[162] D. F. Specht. Probabilistic neural networks. *Neural networks*, 3(1):109–118, 1990.

[163] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[164] A. Storm. Storm, distributed and fault-tolerant realtime computation. 2014.

[165] M. A. Stricker and M. Orengo. Similarity of color images. In *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*, pages 381–392. International Society for Optics and Photonics, 1995.

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

[167] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015.

[168] 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.

[169] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *Systems, Man and Cybernetics, IEEE Transactions on*, 8(6):460–473, 1978.

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

[171] The Apache Software Foundation. Apache lucene - index file formats. https://lucene.apache.org/core/3_0_3/fileformats.html#Definitions. [last visited, Oct., 2016].

[172] 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].

[173] 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].

[174] M. M. Tikir, L. Carrington, E. Strohmaier, and A. Snavely. A genetic algorithms approach to modeling the performance of memory-bound computations. In *Proceedings of the 2007 ACM/IEEE conference on Supercomputing*, page 47. ACM, 2007.

[175] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1958–1970, 2008.

[176] C. Torres, D. Antonioli, and R. D. Odze. Polypoid dysplasia and adenomas in inflammatory bowel disease: a clinical, pathologic, and follow-up study of 89 polyps from 59 patients. *The American journal of surgical pathology*, 22(3):275–284, 1998.

[177] A. Trotman. Learning to rank. *Information Retrieval*, 8(3):359–381, 2005.

[178] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582 –1596, sept. 2010.

[179] M. Van Erp and L. Schomaker. Variants of the borda count method for combining ranked classifier hypotheses. In *Seventh workshop on frontiers in handwriting recognition*. Citeseer, 2000.

[180] B. Van Essen, C. Macaraeg, M. Gokhale, and R. Prenger. Accelerating a random forest classifier: Multi-core, gp-gpu, or fpga? In *Field-Programmable Custom Computing Machines (FCCM), 2012 IEEE 20th Annual International Symposium on*, pages 232–239. IEEE, 2012.

[181] 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 Suppl 3:SE1–8, 2012.

[182] W3. W3 JPEG Standard. http://www.w3.org/Graphics/JPEG/itu-t81.pdf. [last visited, Jul. 12, 2016].

[183] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Computer-aided detection of retroflexion in colonoscopy. In *Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on*, pages 1–6, June 2011.

[184] 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*, 17(1):143–152, 2013.

[185] 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*, 18(4):1379–1389, 2014.

[186] 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.

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

[188] Y.-H. Wei, C.-Y. Yang, T.-W. Kuo, S.-H. Hung, and Y.-H. Chu. Energy-efficient real-time scheduling of multimedia tasks on multi-core processors. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 258–262. ACM, 2010.

[189] T. P. Weldon, W. E. Higgins, and D. F. Dunn. Efficient gabor filter design for texture segmentation. *Pattern Recognition*, 29(12):2005–2015, 1996.

[190] T. White. *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.

[191] B. Winther, M. Riegler, L. Calvet, C. Griwodz, and P. Halvorsen. Why design matters: Crowdsourcing of complex tasks. In *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia*, pages 27–32. ACM, 2015.

[192] 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].

[193] J. Yang, J.-y. Yang, D. Zhang, and J.-f. Lu. Feature fusion: parallel strategy vs. serial strategy. *Pattern Recognition*, 36(6):1369–1381, 2003.

[194] Z. Yang, P. Zhao, G. Wang, and S. Li. Mapping loops of multimedia algorithms for coarse-grained reconfigurable architectures. In *Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on*, pages 608–612. IEEE, 2012.

[195] M. Zaharieva and M. Riegler. Media synchronization and sub-event detection in multi-user image collections. In *Proceedings of the 2nd ACM International Workshop on Human-centered Event Understanding from Multimedia*, pages 13–18. ACM, 2015.

[196] M. Zaharieva, M. Riegler, and M. Del Fabro. Multimodal synchronization of image galleries. In *Proceedings of the MediaEval Workshop 2014*. CEUR Workshop Proceedings, 2014.

[197] X. Zhang, Y. Qu, and L. Xiao. Improving distributed workload performance by sharing both cpu and memory resources. In *Proceedings of the Distributed Computing Systems, 2000, 20th International Conference on*, pages 233–241. IEEE, 2000.

[198] M. Zhou, G. Bao, Y. Geng, B. Alkandari, and X. Li. Polyp detection and radius measurement in small intestine using video capsule endoscopy. In *2014 7th International Conference on Biomedical Engineering and Informatics*, pages 237–241. IEEE, 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.

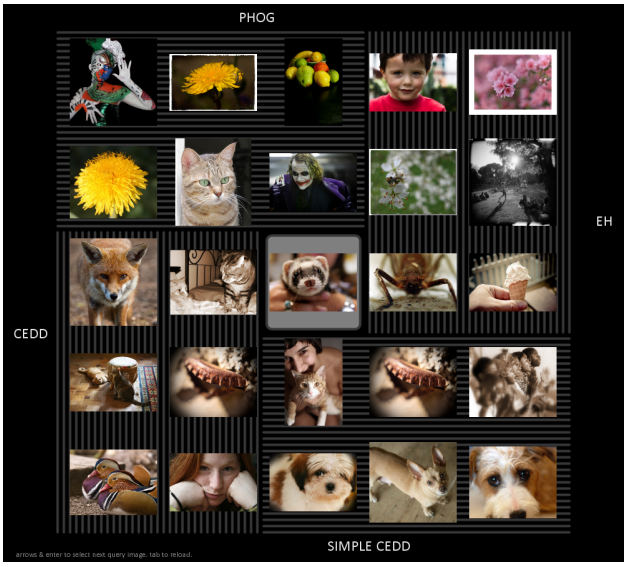



**Figure 1: Sample application built on LIRE. The image in the center is the query, the first six results of four queries based on four different features, three global, one local one, are shown around the query.**

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

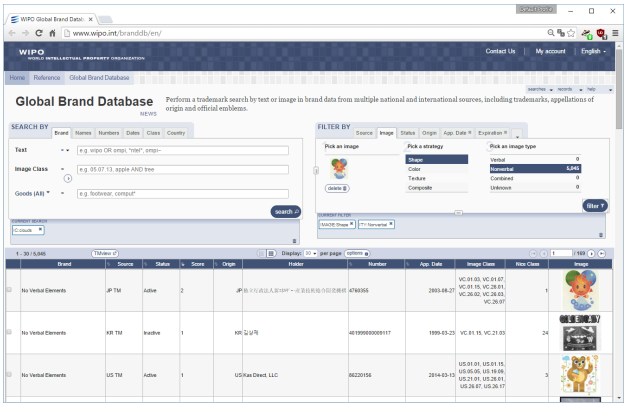



**Figure 2: A screenshot of the UN WIPO Global Brand DB. The image filtering option is implemented using LIRE.**

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.**

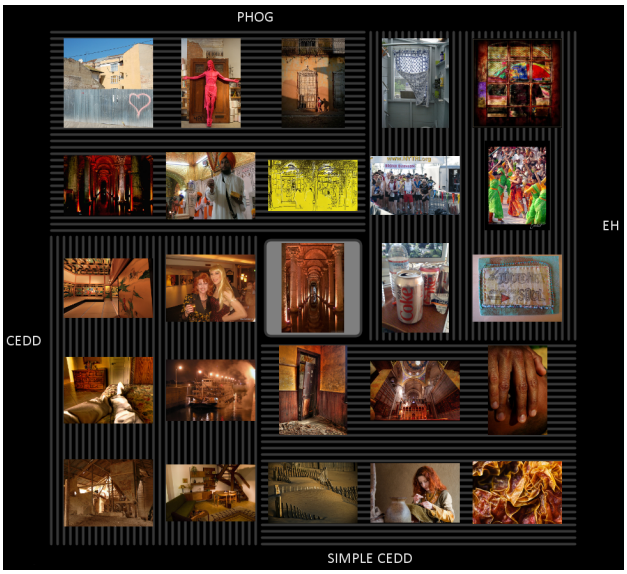



**Figure 3: Sample application built on LIRE showing results for a different query image than Fig. 1.**

## 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

# How 'How Reflects What's What: Content-based Exploitation of How Users Frame Social Images

# How 'How' Reflects What's What: Content-based Exploitation of How Users Frame Social Images

Michael Riegler[2], Martha Larson[3], Mathias Lux[1], Christoph Kofler[3]
[1]Institute for Information Technology, University of Klagenfurt, Austria
[2]Media Performance Group, Simula Research Laboratory AS, Norway
[3]Multimedia Computing Group, Delft University of Technology, Netherlands
michael@simula.no, {m.a.larson, c.kofler}@tudelft.nl, mlux@itec.aau.at

## ABSTRACT

In this paper, we introduce the concept of *intentional framing*, defined as the sum of the choices that a photographer makes on *how* to portray the subject matter of an image. We carry out analysis experiments that demonstrate the existence of a correspondence between image similarity that is calculated automatically on the basis of global feature representations, and image similarity that is perceived by humans at the level of intentional frames. Intentional framing has profound implications: The existence of a fundamental image-interpretation principle that explains the importance of global representations in capturing human-perceived image semantics reaches beyond currently dominant assumptions in multimedia research. The ability of fast global-feature approaches to compete with more 'sophisticated' approaches, which are computationally more complex, is demonstrated using a simple search method (SimSea) to classify a large (2M) collection of social images by tag class. In short, intentional framing provides a principled connection between human interpretations of images and lightweight, fast image processing methods. Moving forward, it is critical that the community explicitly exploits such approaches, as the social image collections that we tackle, continue to grow larger.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## Keywords

Intentional framing; human interpretation of images; user intention; image classification

## 1. INTRODUCTION

Conventionally, multimedia researchers assume that what an image is about is primarily related to its literal subject matter, i.e., the visually depicted entities, events or scenes.


*MM'14,* November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654894.

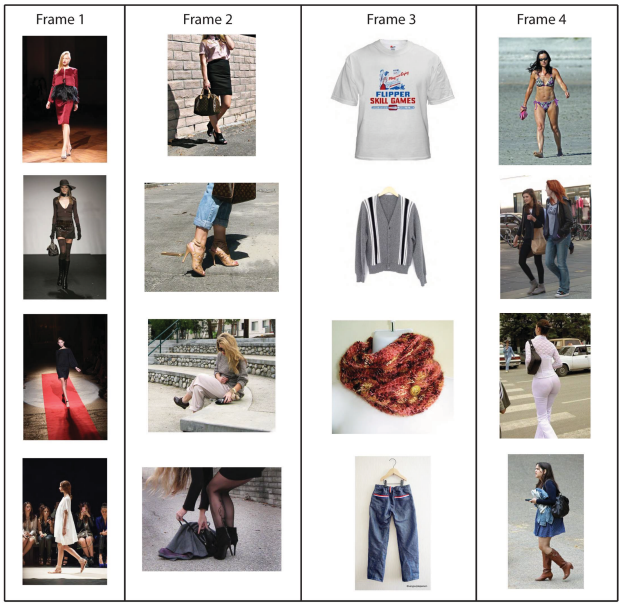



**Figure 1: Four intentional frames reflect four different photographer intents (i.e., *how* users took the images). They are all indicative of the overall topic 'fashion' (i.e., *what* the images depict). Global-feature similarity suffices to capture intentional frames. Computationally intensive approaches are not necessarily required.**

In this paper, we go beyond this conventional viewpoint and demonstrate that *what* an image is about is also reflected in *how* that image was taken. This new perspective benefits content-based approaches to large-scale social image collections, since it can be exploited in a simple, computationally lightweight fashion.

The core of the new perspective is the principle that we refer to as the *intentional framing*. We define intentional framing as, 'the sum of the choices made by photographers on exactly *how* to portray the subject matter that they have decided to photograph.' Note that intentional framing is a photographic act carried out by a photographer. Automatically captured images, such as security camera images, are not expected to exhibit framing effects. Fig. 1 provides an illustration of four ways in which 'fashion' is depicted in user photos on Flickr, a large online social photo-sharing community. These four different cases of *how* the subject matter of a photo is portrayed, correspond to four visually distinctive

*intentional frames*. In Sections 2 and 4, we will return to discuss this figure in more detail, including the relationship between these intentional frames and image similarly calculated automatically on the basis of global image features. Here, we first focus on introducing intentional framing, and describing its importance for multimedia research.

Our definition of intentional framing arises from the following considerations. When taking a photograph, the photographer does not click the shutter randomly, but first decides on a message and a subject. The decision process involves applying, either consciously or unconsciously, a set of conventions. These conventions can be thought as a recipe for a certain kind of image. This recipe is the intentional frame. The fact that the photographer applies a specific intentional frame leads to the generation of an image with distinguishable characteristics. These characteristics are visible in the image, and are used, again, either consciously or unconsciously, by humans in order to interpret the image.

When human viewers interpret an image at the level of its intentional frame, they are making a very high-level semantic judgement. The important role that intentional frame judgements play in human interpretations of images can be illustrated using a short thought experiment. Imagine a home with portraits of family members hanging on the wall. The subject matter (i.e., the *what*) of a portrait image is a person. In taking the image (i.e., the *how*), the photographer had the intent of creating a portrait. What would happen if the portraits of the family members were replaced with mugshots of the family members? This would be a strange situation. A visitor to this home would not easily be able to interpret the wall. A mugshot, like a portrait, portrays a person (i.e., the subject matter has remained the same). However, it is a very different image. The photographer who captures a mugshot has the intention of taking a picture that will be used by the police for identification purposes. This thought experiment demonstrates that two photos with the same literal subject matter (n.B. both a portrait and a mugshot are a photo of a person) are interpreted by the human mind in radically different ways.

The distinction between 'portrait' and 'mugshot' is a simple example used for the purposes of illustration. In this paper, we will investigate neither portraits nor mugshots specifically, but rather use a data-driven approach to explore intentional framing effects in large social image collections. However, by conducting this thought experiment, it is already possible to appreciate the profound implications of the concept of intentional framing for the multimedia field.

First, in order to arrive at image analysis algorithms that are truly capable of mimicking human image interpretation, image analysis algorithms should be 'aware' of the photographer's intention. In other words, they should be able to capture the visual differences that characterize images that were taken with different intents. For example, if humans find the difference between the intentional frames 'mugshot' and 'portrait' to be important, multimedia analysis algorithms need to make this distinction, too.

Second, sensitivity to very high-level semantic judgements, such as those related to intentional frames, will become critical as social image collections continue to grow larger. As pointed out by [13], an image retrieval system that indexes images by detecting basic concepts such as 'dog' cannot effectively support users to search huge social image collections. Even if the relative number of images depicting a 'dog' in the collection is small, if the collection is large enough, a 'dog' detector will detect thousands of dog images, i.e., many more than a user can use in a results' list. Instead, image analysis algorithms are needed which focus on specific aspects of images that are important to users and go beyond the basic concepts they depict. We do not claim that intentional framing is the only way in which human interpretations transcend the literal subject matter of an image. However, it is clear that it is an important contributing factor, and should for this reason be taken into account.

Finally, because of the fact that intentional framing impacts the overall 'look and feel' of images, differences in intentional framing can be captured by simple, lightweight approaches that exploit global representations. Such approaches are critical for allowing image analysis algorithms to scale and handle more and more images, as techniques for image indexing and retrieval are needed for larger and larger collections.

In short, intentional framing is important to the multimedia research community because it provides a principled motivation for applying lightweight approaches, exploiting global-feature representations to large-scale social image collections. The purpose of this paper is to establish the existence of intentional framing as a fundamental principle of human image interpretation, and to demonstrate its importance for content-based approaches to large-scale collections of social images. This paper makes three major contributions: (i) introduce intentional framing as a fundamental principle important for human interpretation of images at a high level of semantic abstraction, (ii) demonstrate that human-perceived similarity with respect to intentional frame corresponds to automatic similarity computed using global-feature representations of images and (iii) show that adopting the intentional framing perspective leads to a back-to-the-basics approach that relies exclusively on global-features to capture image semantics. Our approach delivers image classification rates that compete with the state of the art, while saving significantly in computational complexity.

We finish this section with an overview of the line of argumentation followed in the remainder of the paper. In Section 2, we discuss the related work, and demonstrate that although intentional framing is related to other phenomena studied in the literature, it cannot be reduced to any of them. In Section 3, we explain the concept of intentional framing in greater detail and provide illustrative examples.

Next, Section 4 presents two analysis experiments on human interpretations of images with respect to intentional framing. The experiments involve a user study and explore the judgments that humans make about images at the level of intentional frames. They establish the existence of a correspondence between automatic image similarity calculated on the basis of global features and human perceptions of images with respect to intentional frame.

This correspondence motivates us, in Section 5, to propose a back-to-the-basics simple search approach (SimSea) that leverages global feature representations to classify social images. We report results on standard image-classification task, 2013 Yahoo! Large-scale Flickr-tag Image Classification ACM Multimedia Grand Challenge. The results are surprising: a simple global-feature approach such as SimSea is able to compete with more 'sophisticated' content-based algorithms. Intentional framing, however, constitutes a principled reason why we should actually expect such re-

sults. Our conclusion, presented Section 6, opens up a future perspective.

## 2. RELATED WORK

Our coverage of related work first positions intentional framing with respect to other phenomena related to image semantics. We point to previous work that, plausibly, has taken advantage of the principle of intentional framing, without being aware of its existence. Finally, we cover the lightweight, search-based image classification approaches.

### 2.1 Intentional Frames and Image Semantics

Intentional framing is distinct from other aspects of image semantics because it focuses on 'how' the photographer has realized an image rather than 'what' is depicted in the image. There are three major research areas that seek to analyze images in terms of 'what' they portray: concepts, scenes, and events, which all focus on the literally depicted subject material of an image. Here, we cover each in turn.

*Concepts:* In context of image retrieval and analysis, concepts are objects and other entities that are literally depicted in images. The larger notion of a 'concept' derives from psychology and cognitive science, which has put forth various theories on human concept representations, e.g., concepts as definitions vs. concepts as mental images [33]. Independently of the exact mechanism involved, it is clear that concepts play a role in how humans store, organize and manipulate information about the world around them. For this reason, image analysis research has invested a great deal of effort into developing algorithms capable of detecting visual concepts [10].

An example of the variety covered by concepts is provided by the *ImageCLEF* concept detection task. Here, both categories of high-level semantic abstraction, such as 'fauna', 'age' or 'weather', are used alongside categories of lower abstraction levels, such as 'cat' and 'plant' [1]. Essentially, anything that is nameable by human observer can be considered a concept. Under this perspective, a scene or an event is considered a concept—scenes emphasize the positioning of elements and events include temporal sequence [23]. We turn to discuss both scenes and events in more detail.

*Scenes:* Scene interpretation has its roots in perception psychology. Scene perception describes the visual perception of an environment as seen by an observer at a given time. Rensink [27] describes perception of a scene as high, mid and low level processing steps. Long-term human learning results in a *scene schema* that interlinks the types of objects that occur together.

In the area of machine learning and content-based image retrieval, the notion of *gist* has been used to address the analysis of scenes. Gist originated in language analysis, and was introduced into image analysis by Friedman [7]. In the context of images, the gist is a description of a scene's overall meaning, such as 'farmyard', 'shopping center', or 'city' [27]. Olivia et al. [24] used global features to detect the gist of a scene. Global features capture global attributes of an image related to edges, colors or texture. Hays et al. [8, 9] used the idea of gist to address tasks related to geo-coordinates, such as geo-location detection and geo-scene completion.

The gist of a scene and the intentional framing of an image are related in the way that they both aim to capture global image characteristics. For this reason, global feature representations are suitable for both. However, intentional framing is a much broader notion that gist, since gist is restricted to 'what' is depicted in *scenes*, and intentional framing encompasses 'how' the subject material is presented in a *general social image*. The difference between scenes and intentional frames can be appreciated by considering Fig. 1. The notion of scenes is not adequate to account for the difference between the four frames. Instead we introduce intentional framing to go beyond the gist of scenes and to capture these differences.

*Events:* An event is a specific incident taking place at or over a given time span, involving one or more actors or objects and a specific place. Events often provide subject material for social images: weddings, parties, concerts, and sports events are favorite subjects of photos that users take and share online. Specific to the area of image analysis, an image may depict an event, but it is usually just a snapshot of the event and cannot cover every single aspect of it [12].

Work that has been done on Multimedia Event Detection, exemplified by the work in [22], is devoted to the detection of specific types of human activities, corresponding to types of events. This work focuses on detecting instances of particular event types, e.g., identify multimedia content that depicts a 'kiss' as a human activity. In contrast, a newer breed of work done on Social Event Detection is devoted to detect multimedia content that depicts a specific social event. This work focuses on identifying, for example, whether a photo was captured at a particular wedding. Examples of work on social events include [28], which uses candidate-retrieval methods and machine learning functions to automatically detect events in a stream and [26], which tackles social event detection by using multi-modal clustering and the integration of supervisory signals. Image analysis aiming to identify events, does not cover the same range of phenomena addressed by intentional framing. Note that although human activities and events are depicted in the images in Fig. 1, they do not provide a complete characterization of the differences between the four intentional frames.

To sum up, intentional framing, which focuses on 'how' images are taken, plays a significant role in human image interpretation. This role goes beyond aspects of image semantics that focus on 'what' is depicted in images, including concepts, scenes, and events. We close by mentioning a additional difference between 'the how' and 'the what' of images. Users/viewers recognize that two pictures are similar with respect to an intentional frame—referring again to Fig. 1, note the similarities among the images in each column. This recognition does not imply that it is easy, or even possible, to give an intentional frame a specific name. In contrast, concepts, scenes and events are often readily nameable (e.g., 'cat', 'farmyard' and 'kiss' above). We find that the fact that intentional frames are so difficult to be named, helps to explain why this important principle has been overlooked by the multimedia community thus far. This paper aims to compensate for past inattention.

### 2.2 Covert Exploitation of Framing

Although our basic position is that intentional framing has been overlooked by the multimedia community, we do *not* claim that it has never before been exploited. In this paper, we make the case that intentional framing is an integral part of the act of creating a photo and that, for this reason, we should expect the visual reflexes of intentional framing to act as a social signal that gives rise to exploitable pat-

terns in large collections of social images. If we consider intentional framing to be a fundamental principle underlying human image interpretation, then it is odd to assume that the multimedia community has entirely missed its existence up until this moment. Instead, we consider it to be highly likely that past work in the area of image analysis and retrieval has made use intentional framing effects without being aware of it. Specifically, we make the point that any approach that exploits content based comparison, e.g., pairwise similarity, of images may also be capturing regularities in 'how' the images were photographed alongside of regularities in 'what' the images depict.

Here we mention a two specific social image analysis approaches that we suspect might already exploiting 'how' alongside of 'what'. In Li et al. [15], social tag relevance is learned with a visual neighbor voting algorithm. The approach searches for similar images based on a query image. It cannot be excluded that such similarity is indirectly picking up on 'how' images are taken, in addition to 'what' they literally depict. Another example is Liu et al.'s [16] work on tag propagation. This work makes use of a tag-specific visual sub-vocabulary. Such a sub-vocabulary could easily be exploiting 'how' images are photographed alongside of 'what' they depict. We believe that there are a large number of examples of research that may be unwittingly exploiting intentional framing. An key contribution of this paper is to point out the existence of intentional framing, with the goal of stimulating research on its explicit exploitation. If an algorithm already benefits implicitly from intentional framing, we believe it can only be improved by understanding the extent of this benefit, and by actively seeking to enhance it. In this paper, we do not directly quantify the benefits of intentional framing, but rather focus on laying a solid groundwork for future work in that direction.

## 2.3 Search-based Image Classification

Finally, we turn to discussing work related to our search-based image classification approach, SimSea. We would like to explicitly point out that SimSea itself does not constitute a major contribution of this paper. Rather, we introduce SimSea as a back-to-the-basics algorithm that exploits global-feature representations. Its effectiveness is rather mysterious, until we take the perspective that global features are capable of capturing the semantics of large-scale social image collections because they are sensitive to the semantics associated with intentional frames.

SimSea is a search-based approach representing a variant of the well-known k-NN algorithm. The multimedia item to be classified is used as a query, and a similarity metric is applied to retrieve a ranked list of the most similar items in a collection of multimedia items that has been labeled with category labels. The category labels of the top-ranked items are then propagated to the query image. The work most closely related to ours is the geo-visual ranking approach to content-based prediction of image location [14]. Here, the location of a photo is predicted by using the photo as a query to retrieve a list of geo-visual neighbors from a social image collection, and propagating the most visual likely location to the photo.

Additionally, Wang et al. [34] and Yang and Hanjalic [35] use similar approaches in their work. However, these use both image features and text features, and focus on re-ranking search results. In contrast, our approach relies exclusively on the visual channel, is not deployed for concept detection, and is tested at a much larger scale.

# 3. INTENTIONAL FRAMING

In this section, we present the concept of framing in more detail. Specifically, we discuss photographers' choices that lead to intentional framing, and we provide examples of intentional framing in social image collections, illustrating its link to human interpretations and its generality.

In the most general sense, frames are organizational structures in which information is communicated or understood. They have been extensively studied in the field of communication, which investigates a wide and disparate range of framing phenomena [6, 29]. Across phenomena, however, it is agreed that frames regulate *how* information is communicated, rather than directly determining *what* is communicated. In describing how frames work, Entman [6] states, 'Frames highlight some bits of information about an item that is the subject of a communication, thereby elevating them in salience' (p. 53). Similarly, the decisions that a photographer makes when taking a photograph determine which information in the photographs gets noticed or interpreted as important by the viewer.

## 3.1 Photographers' Choices

Recall that we have defined intentional framing as, 'the sum of the choices made by photographers on exactly *how* to portray the subject matter that they have decided to photograph.' We use the term 'intentional framing', rather than simply 'framing' to emphasize that the 'frame' is the visible reflexes of the *intent* of the photographer to create a certain type of image. The term 'intentional framing' also disambiguates our use of the word 'framing' from another use common in photography. Specifically, photographers use 'framing' to refer to positioning the subject of a photo within a door or other opening that acts like a window frame in a photograph. This sense of 'framing' is not the one that we are addressing here.

Choices a photographer makes to achieve certain types of framing include color distribution, lighting, positions of objects and people, camera angle, depth of field, and focus. They also include the choice of the precise moment during ongoing action at which the image is shot. In this way, the photographer also influences exactly what is depicted in the image, for example, facial expressions of the people appearing in the image. In general, the influence of the photographer reflects not so much personal choices, but rather shared expectations between photographers and viewers about how photos portray the world. These expectations constitute a set of conventions that allow viewers to interpret photos. Radically creative photography may make breaks with conventions, but photos that stray too far from familiar recipes are difficult to interpret.

The importance of intentional framing for photographers is witnessed by the way how it is described on websites that teach photography. For example, Fodors provides a web tutorial for travel photography [31]. Several different methods for framing photos are described, each related to different subject matter: 'classic vacation shots', 'the man-made world', 'the natural world', 'the elements' and 'people'. Each is broken down into finer-grained topic related categories.

Clearly, the decisions that photographers make that determine intentional framing are closely related to composi-

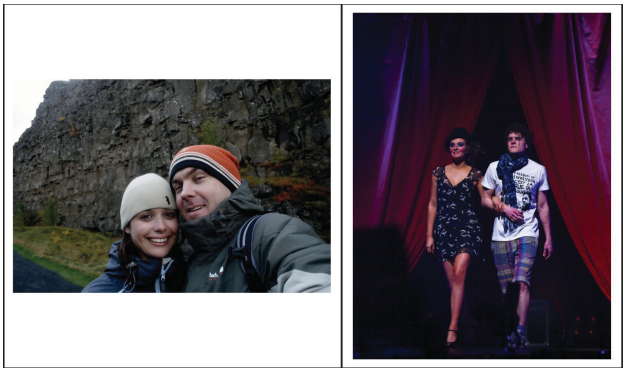


**Figure 2: Example Flickr images that depict the same visual contents (a woman and a man), but correspond to two different intentional frames (left: holiday memories and right: fashion)**

**Table 1: The metadata for the images in Fig. 2 reflecting the different underlying intents of the user/photographer**

| | Left image | Right image |
|---|---|---|
| **Title** | Thingvellir | DCU Fashion Show 2009 |
| **Tags** | trip, iceland, reykjavik, ... | fashion, Cirque Du Couture dcu, Couture, ... |
| **Description** | Iceland 2009. | DCU Style Society presents DCU FASHION SHOW 2009 ... |

tion. The composition of a photograph includes the arrangement of objects, the angle, the focus or the distribution of colors in a photograph. Although composition choices contribute to intentional framing, intentional framing cannot be reduced to composition. We explicitly point out that intentional framing also includes choices beyond composition, such as whether human subjects are visibly expressing emotion, and the choice of the exact setting. The fact that photographers consider intentional framing as a way to extend beyond composition is illustrated by the organization of the tutorial [31]. Here, methods for framing photos are *not* treated under the heading 'Photography composition rules'. There are rather separate sections dedicated to composition and to framing. We are interested in the broader concept of framing rather than the narrower concept of composition as it is more tightly related to the topic of the image, which makes it a better indicator of image semantics.

## 3.2 Intentional Frames in Practice

In Fig. 2, we present two social images from Flickr that both depict the same basic content, a woman and a man, but differ in respect to their intentional framing. We can gain insight into the intent of the users that took these images by inspecting their titles, tags and descriptions, shown in Tab. 1. This metadata leads to the conclusion that the intent behind the image on the left is to capture a memory of a trip and the intent behind the image on the right is to depict fashion.

This difference in intent can also be seen in how the users who took the photos have chose to frame them. Although both images show a man and a woman, in the image on the left, the user has chose to make a 'selfie' in an outdoor setting that focuses on faces and smiles, and in the image on the right, the setting is an illuminated stage and the focus is on the clothing. An inspection of Fig. 2 reveals that these choices have visual reflexes in the photos. The visual manifestations of intentional framing signal to the viewer that one photo should be interpreted as representing holiday memories and the other as depicting fashion.

With this example we would also like to stress the point that textual metadata could possibly help in the differentiation of photos on the basis of their intentional frame. Our focus here, however, is on visually observable intentional framing effects and on content-based approaches. For this reason, we do not consider textual features any further.

## 3.3 Viewers' Interpretations

The study of framing has its roots in the field of social psychology, where a frame describes a general, mainly subliminal, basic idea at play during perception or interpretation. The notion of 'frame' is thus tightly related to *Gestalt*, the perception of the essence or shape of an entity's complete form [11]. Specific to image perception is the notion of 'gist' [7], i.e., what is perceived from an image at a glance. We have already noted that gist-based methods have been applied by multimedia researchers to the problem of analyzing images that depict scenes.

Viewer interpretations of images are tightly synchronized with the intentional framing that is chosen by a photographer. In fact, the intentional framing of the image constitutes a signal from the photographer to the viewer about how the image should be interpreted. For some subject material, photographers often use highly conventionalized intentional frames. For example, nearly everyone can bring a standard picture in mind of how a traditional bridal pair appears in a wedding photo, or a how a public figure is represented in a certain role, e.g., a politician delivering an inspiring speech.

For other subject material, the intentional framing is less tightly linked to the subject matter, but rather more closely related to the underlying goal or purpose. For example, [18] establishes a typology of photographer intentions. This work demonstrates the reasons for which people take social images range from sharing emotions to recalling a feeling or collecting and storing information.

Our work does not depend on explicitly identifying or cataloguing intentional frames corresponding to all possible image topics, or photographer goals and purposes. We are rather interested in the fact that photographers use intentional frames to create photos, and that users/viewers differentiate photos on the basis of intentional frames. In other words, our work is focused on establishing that, alongside of *what* photos depict, *how* photos are taken is important for human interpretation of image semantics.

We point out that intentional framing is closely linked to the notion of *connotation*. In the area of images, connotation refers to those aspects of image interpretation that go beyond the literally depicted subject material of the image. In his seminal essay in [2], *The Photographic Message*, Roland Barthes characterizes connotation as 'the imposition of second meaning on the photographic proper' (p. 20). Intentional framing can also be considered a 'second meaning'. However, understanding connotation involves interpreting 'what' is depicted in an image. For example, red roses are commonly considered to have connotations of love. In contrast, intentional framing keeps the focus specifically on 'how' image content is depicted, and the way that photog-

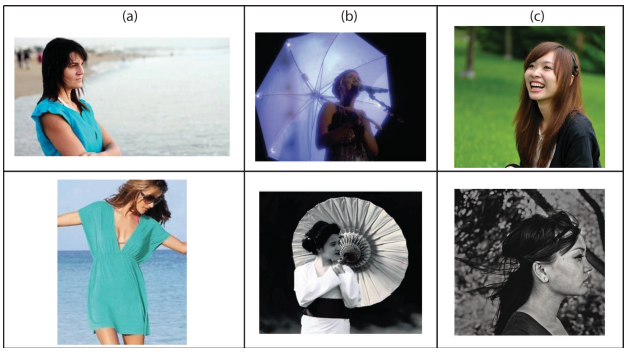



**Figure 3: Pairs of photos that contrast with respect to intentional frame. The pairs differ with respect to the interpretations of human viewers, impressionistically described as: (a) feeling vs. fashion, (b) performance vs. history (c) personality vs. art**

rapher choices related to 'how' are reflected in visual characteristics of an image. We remark that ultimately intentional framing may lead the multimedia research community to more effective exploitation of overall image connotations.

## 3.4 Generality of Intentional Framing

Intentional framing is a general phenomenon underlying social images. The visual reflexes of photographers' intent constitute a social signal that influences the global patterns that exist in a large collection of social images. Some photographers might take images unthinkingly, but most conceptualize their images to at least a minimum extent. Photographer choices, in turn, impact exactly *how* the subject material is depicted in the image. We do not claim that intentional framing constitutes a strong signal within a social image collection. Rather, our position is that this signal exists, and that it is strong enough to be effectively exploited. Here, we present additional examples to demonstrate that intentional framing takes different forms, and that a large range of images can be differentiated on the basis of intentional framing.

The Flickr images in Fig. 3 are arranged in pairs that differ with respect to intentional framing. The contrast between the two photos in each pair demonstrates that if two photos depicting the same concept or entity use different intentional frames, the result is two images with different interpretations. Consider the two photos in column (a). The top photo is about what the woman in the blue dress is feeling, and the bottom photo is about the blue dress. The contrast is due to the framing choices made by the photographers who took these photos. These choices include not only the ratio of the frame filled by the dress vs. the ratio filled by the background water, but also with the depth of field, the overall color palette, and the emotion projected by the human subject, and the subject's posture. In other words, it is intentional framing and not the depicted visual concepts that serve to distinguish these images from the point of view of a human interpreter.

Similar observations can be made about the photo pairs in (b) and (c). In (b), one set of photographer choices leads to an image depicting an ongoing performance (top), and the other to an image that documents history (bottom). Note that these two photos are very similar with respect to their basic composition, but different in their interpretation. This pair illustrates how intentional framing includes, but goes beyond, photographers' composition choices. In (c), one set of photographer choices leads to an image that conveys the happy personality of the subject (top) and another set of choices lead to a photo with a somber mood (bottom) that could be considered a work of art, more than a testimony to the personality of the person displayed.

It is important to note that the descriptions we use to refer to viewer interpretations of framing are impressionistic. We do *not* claim that these are the only possible descriptions, or that algorithms should predict these interpretations directly. Our point is that intentional framing is important for human image interpretations, and content-based algorithms should not be 'blind' to its existence. The larger message is that multimedia researchers should not indiscriminately assume that content-based image methods must 'compensate' for the visual variability of depicted objects, scenes, and images. Such approaches will lead to image analysis and retrieval systems that cannot possibly be sensitive to the difference in human interpretation between the pair of images in (a). Instead we advocate systems that admit the possibility that differences important for human semantic interpretation of images are related to intentional frames.

# 4. HUMAN VIEWS ON FRAMING

In this section, we empirically investigate the phenomenon of intentional framing. On the basis of the discussion in Section 3, we expect intentional framing to manifest itself in a collection of social images in the form of clusters of images that are homogenous in terms of their overall 'look and feel'. For this reason, we study clusters of images that are created automatically using global feature representations. We are interested in two aspects of these clusters, which we investigate in two analysis experiments involving human judgments collected via user studies.

The first experiment explores the correspondence of global-feature clusters with human judgements of photographer intent. The second experiment explores the correspondence of global-feature clusters with image semantics in the form of a higher level topic, in this case, 'fashion'. Each experiment consists of two steps, first, the *clustering step*, in which we create clusters in a social image collection, and, second, the *correlation step* in which we analyze the relationship between the clusters and human judgements related to intentional framing.

## 4.1 Global Features and Photographer Intent

According to the principle of intentional framing, the intent of the photographer guides the decisions made by the photographer while conceptualizing an image, resulting in an image with a particular intentional frame. However, since intentional framing results from a general recipe for a photograph, rather than specific rules, and, since photographers apply this recipe only to varying degrees, we, yet, know nothing about the visual variability that characterizes intentional frames. For this reason, the goal of our first experiment, is to demonstrate that it is indeed conceivable that global features can capture the regularities of frames.

For this experiment, we use the *Photo Intentions* data set that has been created by Lux et al. [20], and consists of 1,310 Flickr photos annotated with *photographer intent categories*. The categories correspond to general photographer goals in taking a picture: (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*, and were chosen on the basis of a previous user study carried out by [18]. The images in the data set were annotated by the users who took them, who were contacted by Lux et al. [20] via Flickr. The category labels provided by the photographers were verified using a crowdsourcing experiment carried out on Mechanical Turk. As explained in detail in [20], five crowdworkers judged each image, and rated it with respect to each of the six intent categories. These ratings, used in our experiment, reflected the association of the image with each of the six categories using a 5-point Likert scale.

It is important to note that in this experiment, we do not assume that the photographer's intent category corresponds just to one single intentional frame. Instead, we take these categories to involve multiple closely related intentional frames that photographers use to accomplish a particular goal or purpose. We assume that if visual clusters correspond to intentional frames, then they will also be correlated with intent categories that encompass multiple frames. To our knowledge, our data set is the largest publicly available collection of social images that includes information about the intent of the photographer.

The *clustering step* in our experiment was carried out as follows. A selection of common global features was made, and the features were extracted from the images using *LIRE* (latest version[1]) [19]. For each type of global feature, clustering is performed using *Weka* (version 3[2]). We chose X-means clustering [25], since it determines the number of the clusters automatically, which is important for the experiment.

The *correlation step* was carried out for each different global-featuring clustering of the images. The purpose of the correlation step was to compare the visual closeness of the images in a cluster, with the human perception of whether the images were 'close' with respect to the intent of the photographer. We analyze each global feature clustering with respect to each intent category separately. Specifically, we calculate the Pearson correlation between the mean square distance of the images in a cluster from their cluster centroid and the mean of the Likert-scale ratings reflecting the degree to which the images in the cluster are associated with the intent category.

Tab. 2 summarizes the results, and demonstrates that the experiment uncovered the existence of a number of cases in which the tightness of visual clusters correlates ($> 0.3$) with human agreement on the photographer's intent.

These cases (n.B. they are negative correlations) are indicated with black. It can be seen that certain features seem to be particularly well suited for certain categories. For example, FCTH is the best feature for detecting photos for which the photographer's intent was *publish on-line* (i.e., in a blog). It is important to note that the results of this experiment must be seen as a *suggestion* that global features can capture photographer intent. There are also cases of positive correlation, which are marked in white, where it is clear that other effects are at play. However, if there was no relationship between global features and photographer intent, we would have expected a table that was entirely grey,

[1]https://code.google.com/p/lire/source/checkout
[2]http://www.cs.waikato.ac.nz/ml/weka/downloading.html

**Table 2: Correlation of global-feature-based clusters (MSE) and human agreement on photographer intent on the *Photo Intentions* data set.**

| Feature | recall situation | preserve good feeling | publish online | show to family and friends | support task | preserve bad feeling |
|---|---|---|---|---|---|---|
| CEDD | | | | | | |
| FCTH | | | | | | |
| Gabor | | | | | | |
| Tamura | | | | | | |
| Luminance Layout | | | | | | |
| Scaleable Color | | | | | | |
| Opponent Histogram | | | | | | |
| AutoColor Correlogram | | | | | | |
| JPEG Coefficent | | | | | | |
| Edge Histogram | | | | | | |
| PHOG | | | | | | |
| JCD | | | | | | |
| JointHistogram | | | | | | |

which is not the case. Encouraged by this initial experiment, we turned to a second, larger-scale experiment, that investigates the connection between image clusters and topical semantics.

## 4.2 Intentional Framing and Topic

Our second analysis experiment is closely related to the title of this paper. It investigates the connection between 'how' an image is taken and 'what' that image depicts. Recall that the intentional frame that a photographer chooses is related to the particular subject material that is portrayed in the image (i.e., the topic). The importance of the relationship between intentional framing and topic is the following: if visual patterns of intentional framing in social data sets are indicative of topic, then they can be exploited for content-based tasks such as analysis of image semantics, and image retrieval.

For this experiment we use the *Fashion 10000* data set for Flickr images, which was created by Loni et al. [17] for the purpose of developing classifiers to detect fashion images in social image collections. The data set consists of 30,000 images and was collected to contain a significant portion of images (>10,000 of them) that are related to fashion and clothing. Further details on how the +fashion/-fashion labels were generated can be found in [17].

On the basis of the results of the previous experiment, we expect that clustering using global feature representations are indeed capable of picking up visual regularities in the data related to intentional framing. This experiment had the goal of uncovering a relationship between the visual tightness of clusters and human judgements that these clusters were related to the overall topic of fashion.

Because the *Fashion 10000* data set is an order of magnitude larger than the intention data set, we first carried out clustering, and then submitted the clusters to a group of human judges. The *clustering step* in this experiment was carried out by first determining an optimal global feature representation for the data using average information gain. Under this assumption, the following features were identified as useful for the data set: CEDD, FCTH, JCD, PHOG, ColorLayout, JPEG coefficient histogram and ScalableColor [19]. As before, X-means clustering was applied, resulting in 62 clusters.

In the *correlation step*, a set of human subjects were presented with 62 screens of images, each screen containing im-

ages sampled from one of the 62 clusters (n.B. the clusters were too large to judge in their entirety). In total, there were 10 participants who judged the images. The participants were selected by convenience sampling from the immediate environment of the authors, and briefed to ensure that they had an adequate understanding of social fashion. The participants then judged the consistency of the clusters with respect to fashion.

As in the previous experiment, we calculate the Pearson correlation between the mean square distance of the images in a cluster from the cluster's centroid (reflecting visual tightness of the cluster) and the average human agreement about the fact that the cluster reflects fashion. The result was a negative correlation, $-0.56$ (i.e., small, tight clusters are associated with clear human judgements of topic focus). This result provides evidence that global features are indeed able to build clusters that partition fashion from non-fashion images in the data set. This experiment supports the conclusion that 'how' images are taken, as reflected in global-feature-based clustering, is indicative of the topical subject matter that they contain.

We conclude this section by mentioning that qualitatively the outcomes of the second experiment were striking. Specifically, the four groups of intentional frames in Fig. 1 in Section 1 were *not* hand selected from the data. Rather, these four intentional frames represent clusters that were formed using global-feature-based clustering in the second analysis experiment. These clusters serve as a compelling illustration of the link between global features and semantic image content. Our position is that this link exists, because of the photographer's tendency to take pictures of specific content which follows a set of intentional frames. This effect is stable enough to be a useful visual signal within large-scale social image collections.

# 5. CONTENT-BASED CLASSIFICATION

The evidence in Section 4 suggest that there is a link between global feature representations and image topic that is mediated by intentional framing. Motivated by this evidence, we carry out an experiment designed to exploit that link. The experiment involves large-scale classification of social images into tag-classes. Intentional framing gives us reason to believe that images belonging to a certain tag-class, and therefore containing certain topical subject material, will be characterized by patterns of intentional framing. These patterns reflect typical sets of choices made by users on 'how' to make a photograph that are related to the subject material that they are photographing. We do not expect the effects to be strong. Instead, our goal is to present plausible proof that the effects of intentional framing are exploitable for a task related to image semantics.

## 5.1 Data Set and Experimental Setup

We carry out our content-based classification experiment on the *Yahoo! Flickr Creative Common Images tagged with ten concepts, version 1.0 data set*[3] that was used for the 2013 Yahoo! Large-scale Flickr-tag Image Classification ACM Multimedia Grand Challenge. The data set consists of 1.5 million training images associated with ten equally-sized tag-classes and 500,000 test images. The tag-classes are: *2012*, *beach*, *food*, *london*, *music*, *nature*, *people*, *sky*, *travel*, and *wedding*. This data set is considered challenging not only due to its scale, but also because each topical tag-class is characterized by a very high degree of visual variability. Our choice of a standard data set allows us to compare our approach to the performance achieved by current state-of-the-art methods.

In order to make clear how the theory of intentional framing is expected to contribute to the performance of a classifier on such a classification task, we consider the class 'London' in more detail. Images taken all over London will be tagged 'London', giving rise to a high level of visual diversity. However, because we are looking specifically at social images, we expect that people are taking pictures of London mainly with the intention of documenting the city, for example, as tourists, as residents or as journalists. For this reason, we expect photos to be generally associated with key intentional frames, examples might be, cityscape photos, photos that emphasize a sense of place, and photos taken to preserve memories. These intentional frames are indicative of the topic 'London' the way that the four frames in Fig. 1 are indicative of the topic 'fashion'.

The intentional frames are not expected to be mutually exclusive among tag-classes. However, they are expected to support discrimination well enough to act as indicators of tag-classes. For example, it is reasonable to expect that more cityscape photos would be anticipated in the tag-class 'London' than in the tag-class 'Food'. In this way, intentional framing can be anticipated to deliver performance on this task—sensitivity to the specific literally depicted content of the images (i.e., detecting specific food items or specific city landmarks) is not necessary.

Our simple search classification approach, SimSea, is a variant of the k-NN algorithm. We choose a back-to-the-basics approach because of its computational simplicity and the speed that it delivers on large scale image classification problems. As previously mentioned, SimSea is not itself novel. Our novel contribution is that intentional framing provides a principled explanation as to why an algorithm such as SimSea should work well on a large collection of social images.

SimSea is implemented by extending the LIRE framework [19] with an implementation of a search based classifier. The following global features were used: JCD [5], CL [4], OH [32] and PHOG [3]. These features were selected using information gain calculated on the development set (a subset of the training set).

For retrieval we employ the inverted index strategy to index the hash values, like terms describing the actual image. To query the system, we create a term-based query from the hashes of the query image.

Classification proceeds as follows. Each search result of a given tag-class is counted as one vote in favor of assigning the query image to that tag-class. The class with the most votes wins, and is returned as the classified class for the query image. In case of a draw, the occurrences of the class is weighted by the rank of each image with the same class. The weight function is defined as:

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

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

The class with the highest $ClassScore$ of all classes $c \in C$ is chosen as class $c$ of the image. The $ClassScore$ is calculated

[3]see http://webscope.sandbox.yahoo.com

**Table 3: $iAP$ per class on the development set.**

| | JCD | CL | OH | PHOG |
|---|---|---|---|---|
| 2012 | 0.198 | 0.128 | 0.13 | 0.104 |
| beach | 0.448 | 0.487 | 0.342 | 0.534 |
| food | 0.531 | 0.492 | 0.389 | 0.352 |
| london | 0.244 | 0.201 | 0.146 | 0.347 |
| music | 0.526 | 0.457 | 0.495 | 0.164 |
| nature | 0.502 | 0.41 | 0.435 | 0.503 |
| people | 0.264 | 0.227 | 0.244 | 0.105 |
| sky | 0.628 | 0.601 | 0.544 | 0.473 |
| travel | 0.139 | 0.101 | 0.128 | 0.112 |
| wedding | 0.463 | 0.272 | 0.262 | 0.235 |

by counting the occurrences of each class $c$ and multiply it with the summed $WeightedRankScore$. $rank(I_i) : \{I_i\} \rightarrow \mathbb{N}$ gives the rank of an image. The $WeightedRankScore$ is the sum of $rank(I_i)^{-1}$ scores for each class. The search time of this approach is well below 300 ms for the 1,500,000 indexed images. Due to the nature of the global feature search, the search time will scale sub-linearly with the number of images in the index.

## 5.2 Experimental Results

For evaluation metric, we adopt mean interpolated averaged precision ($MiAP$), which was also used to evaluate the results of the Grand Challenge.

Our first step is to use the development set to determine optimal settings. Our development set is based on a sub-set of the training data that includes 1,000 randomly selected images per class. A setting of $k = 50$, was determined to be optimal. Tab. 3 reports results for $k = 50$ using a variety of global features. It can be seen that average precision is very similar for all choices of features. However, JCD achieves the highest overall mean interpolated average precision (0.417), and is therefore chosen for the experiment on the test set.

Our second step is to carry out classification on whole test data set (500k images; 50k images per tag-class). Applying the optimal settings determined on the development set, i.e., $k = 50$ and JCD features, we performed our second experiment with the test data set featuring 50k images per class which leads to a $MiAP$ over all classes of 0.391.

Tab. 4 shows the $MiAP$ over all classes for the 500k test set and all 1.5 million training images to train the model, compared with the best results of the ACMMM Grand Challenge 2013 from Mantziou et al. [21]. The best performing reported result from the second approach from Su et al. [30] is not compared because the reported $MiAP$ excludes the tag-class *2012*. However, they also use a concept detection approach, Hessian affine (Concept 1 (HA)), which is more comparable. The comparison shows that our *SimSea* approach which uses *only one* global feature, nearly can reach the performance of the other approaches and in one case, Concept 1 (HA) [30], can reach better performance.

As baseline we use the results of SMaL [21] (Local 1 (SMaL)) and SVM [21] (Local 2 (SVM)), which both rely on local features and complex learning algorithms, because they report the $MiAP$ for each class which makes it better comparable with our results. Based on the $MiAP$ We calculated the statistical significance (Wilcoxon Signed-Rank Test) with a significance level of 0.01. This leads to p-value of 0.5754 for SimSea versus Local 1 (SMaL) and a p-value of 0.3320 for SimSea versus Local 2 (SVM). This test shows that the difference is not statistically significant and

**Table 4: SimSea vs. best results from the ACMMM Grand Challenge 2013. The difference with Local 1 and Local 2 is not statistically significant**

| | SimSea | Local 1 (SMaL) [21] | Local 2 (SVM) [21] | Concept 1 (HA) [30] |
|---|---|---|---|---|
| $MiAP$ | 0.391 | 0.422 | 0.413 | 0.37 |

therefore our methods performance cannot be interpreted as worse (or better) than the Local 1 and Local 2 approach. For the sake of completeness we mention that all approaches outperform the dominant class baseline, which is 0.1.

It is important to point out that the run-time performance of our approach is very good. Classification of a single image takes roughly 300 ms on a current Windows 7, Intel Core i7, 16GB PC. This is faster than 10 minutes for Local 1 (SMaL) and 2.5 seconds for Local 2 (SVM) on a 24-core Intel Xeon Q6600 2.0Ghz with 128GB RAM reported in [21].

# 6. CONCLUSION AND OUTLOOK

This paper has presented a proof for the importance of visual patterns in large collections of social images that exist due to underlying consistencies in how photographers choose to make images. Conceptually, we have turned from considering an image as something that is viewed by a user ('what' is shown in the image) to considering an image to be something that was created by a photographer ('how' the image was captured by the photographer). This shift of perspective allows us open up a new set of commonalities between images. These commonalities are useful for image analysis and retrieval because they both connect images at the level of pixels, and also correspond to connections perceived by humans interpreting images.

Specifically, this paper has introduced intentional framing, a principle that accounts for the connection between the decisions made by photographers that are related to the subject material they photograph, and visual characteristics of social images. We show that global feature representations of images are connected to human perceptions of photographer intent, and also that intentional frames chosen by photographers are connected to the semantics of images at the level of topic

We report the results of a large-scale image classification experiment that makes use of a back-to-the-basics simple search approach exploiting global feature representations of images. If we consider that the perspective image topic is exclusively related to 'what' images predict, the good performance of this simple approach is mysterious. Global features are known to be related to the overall 'look and feel' of images and not necessarily to specific semantic content. However, once we understand that intentional framing has the ability to act as a bridge between global characteristics of an image, and interpretations of image semantics, these results are expected.

The principled account that intentional framing provides for the results of our content-based classification experiment opens a new vista for multimedia research. We consider the results of this experiment to reflect the force of intentional framing at work in a large-scale social image collection. However, it measures that force, at best, only indirectly. Additional work is required to understand the exact nature of that force, the factors that influence it, and how it can best be harnessed in the service of image analysis and retrieval. For example, we would not expect an intentional-

framing-sensitive approach to work well on a collection of images not taken by human photographers, e.g., Google street view images. Such a collection would have only a single robotically created frame, and image semantics would not be differentiated at this level. Further experimentation is necessary to test this hypothesis and to discover how to exploit the intentional framing principle to its full potential.

Taken as a whole, the evidence presented in this paper serves to demonstrate the high promise of the 'how' perspective for social images. We conclude that intentional framing opens the possibility of the exploitation of lightweight approaches that contribute to the development of a new breed of fast image analysis and content-based retrieval algorithms for large-scale social image collections.

## 7. ACKNOWLEDGMENTS

We would like to thank Xinchao Li, Yue Shi and Oge Marques for fruitful discussion of framing. M. Riegler is a recipient of the Excellence Scholarship of the Industrialists' association Carinthia. which provided partial support for this research. Support was also received from Lakeside Labs GmbH, Klagenfurt, Austria, the European Regional Development Fund, the Carinthian Economic Promotion Fund (KWF-20214/25557/37319), EC FP7 project CUbRIK (287704), the iAD center for Research-based Innovation (174867) funded by the Norwegian Research Council, and the Dutch national program COMMIT. C. Kofler is a recipient of the Google Europe Doctoral Fellowship in Video Search, which also provided partial support.

## 8. REFERENCES

[1] ImageCLEF Task. http://imageclef.org/2012. [lv., 08, 13].

[2] R. Barthes. *Image Music Text*. Hill and Wang, 1977.

[3] A. Bosch, A. Zisserman, and X. Munoz. Representing Shape with a Spatial Pyramid Kernel. In *Proceedings of the CIVR '07*, pages 401–408, New York, NY, USA, 2007.

[4] 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.

[5] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux. Selection of the Proper Compact Composite Descriptor for Improving Content Based Image Retrieval. In *Proceedings of the SPPRA 09'*, 2009.

[6] R. M. Entman. Framing: Toward Clarification of a Fractured Paradigm. *Journal of communication*, 43(4):51–58, 1993.

[7] A. Friedman. Framing Pictures: The Role of Knowledge in Automatized Encoding and Memory for Gist. *Journal of experimental psychology: General*, 108(3):316, 1979.

[8] J. Hays and A. A. Efros. IM2GPS: Estimating Geographic Information from a Single Image. In *Proceedings of the CVPR 08'*, pages 1–8.

[9] J. Hays and A. A. Efros. Scene Completion Using Millions of Photographs. In *Proceedings of the ACM TOG 07'*, volume 26, page 4, 2007.

[10] M. J. Huiskes, B. Thomee, and M. S. Lew. New Trends and Ideas in Visual Concept Detection: The MIR Flickr Retrieval Evaluation Initiative. In *Proceedings of the ACM ICMR 10'*, pages 527–536, 2010.

[11] G. Humphrey. The Psychology of the Gestalt. *Journal of Educational Psychology*, 15(7):401, 1924.

[12] J. Kim. Events as Property Exemplifications. In *Action theory*, pages 159–177. Springer, 1976.

[13] M. Larson, M. Melenhorst, M. Menéndez, and P. Xu. Using Crowdsourcing to Capture Complexity in Human Interpretations of Multimedia Content. In *Fusion in Computer Vision*, pages 229–269. Springer, 2014.

[14] X. Li, M. Larson, and A. Hanjalic. Geo-visual Ranking for Location Prediction of Social Images. In *Proceedings of ACM ICMR 13'*, pages 81–88. ACM, 2013.

[15] X. Li, C. G. Snoek, and M. Worring. Learning Social Tag Relevance by Neighbor Voting. *Multimedia, IEEE Transactions on*, 11(7):1310–1322, 2009.

[16] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang. Image Retagging Using Collaborative Tag Propagation. *Multimedia, IEEE Transactions on*, 13(4):702–712, 2011.

[17] B. Loni, L. Y. Cheung, M. Riegler, A. Bozzon, L. Gottlieb, and M. Larson. Fashion 10000: An Enriched Social Image Dataset for Fashion and Clothing. In *Proceedings of ACM MMSys 14'*, pages 41–46, New York, NY, USA, 2014. ACM.

[18] M. Lux, M. Kogler, and M. del Fabro. Why Did You Take This Photo: A Study on User Intentions in Digital Photo Productions. In *Proceedings of SAPMIA 10'*, SAPMIA '10, pages 41–44, New York, NY, USA, 2010. ACM.

[19] M. Lux and O. Marques. Visual Information Retrieval Using Java and LIRE. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 5(1):1–112, 2013.

[20] M. Lux, M. Taschwer, and O. Marques. A Closer Look at Photographers' Intentions: A Test Dataset. In *Proceedings of the ACM CrowdMM 12'*, pages 17–18. ACM, 2012.

[21] E. Mantziou, S. Papadopoulos, and Y. Kompatsiaris. Scalable Training with Approximate Incremental Laplacian Eigenmaps and PCA. In *Proceedings of the ACM MM 13'*, pages 381–384, 2013.

[22] T. B. Moeslund, O. Javed, Y.-G. Jiang, and R. Manmatha. Special Issue on Multimedia Event Detection. *Mach. Vision Appl.*, 25(1):1–4, Jan. 2014.

[23] M. Naphade, J. R. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale Concept Ontology for Multimedia. *MM IEEE*, pages 86–91, 2006.

[24] A. Oliva and A. Torralba. Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Progress in brain research*, 155:23–36, 2006.

[25] D. Pelleg, A. W. Moore, et al. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of ICML 00'*, pages 727–734, 2000.

[26] G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Social Event Detection Using Multimodal Clustering and Integrating Supervisory Signals. In *Proceedings of ACM ICMR 12'*, page 23, 2012.

[27] R. A. Rensink. Scene Perception. 7:151–155, 2000.

[28] T. Reuter and P. Cimiano. Event-based Classification of Social Media Streams. In *Proceedings of ACM ICMR 12'*, page 22, 2012.

[29] D. A. Scheufele. Framing as a Theory of Media Effects. *Journal of communication*, 49(1):103–122, 1999.

[30] Y.-C. Su, T.-H. Chiu, G.-L. Wu, C.-Y. Yeh, F. Wu, and W. Hsu. Flickr-tag Prediction Using Multi-modal Fusion and Meta Information. In *Proceedings of ACM MM 13'*, pages 353–356, 2013.

[31] F. Travel. Framing of Images from Photographers View. http://www.fodors.com/travel-photography/. [lv., 11, 13].

[32] K. E. van de Sande, T. Gevers, and C. G. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596, 2010.

[33] S. Vinner. Concept Definition, Concept Image and the Notion of Function. *International Journal of Mathematical Education in Science and Technology*, 14(3):293–305, 1983.

[34] L. Wang, L. Yang, and X. Tian. Query Aware Visual Similarity Propagation for Image Search Reranking. In *Proceedings of ACM MM 09'*, pages 725–728. ACM, 2009.

[35] L. Yang and A. Hanjalic. Supervised Reranking for Web Image Search. In *Proceedings of ACM MM 10'*, pages 183–192. ACM, 2010.

# Paper III

# Exploitation of Producer Intent in Relation to Bandwidth and QoE for Online Video Streaming Services

# Exploitation of Producer Intent in Relation to Bandwidth and QoE for Online Video Streaming Services

Michael Riegler[1], Lilian Calvet[1], Amandine Calvet, Pål Halvorsen[1], Carsten Griwodz[1]
[1]Media Performance Group, Simula Research Laboratory & University of Oslo, Norway
{michael, paal, lcalvet, griff}@simula.no, amandine.calvet@gmail.com

## ABSTRACT

This paper is the product of recent advances in research on users' intent during multimedia content retrieval. Our goal is to save bandwidth while streaming video clips from a browsable on-demand service, while maintaining or even improving the users' quality of experience (QoE). Understanding user intent allows us to predict whether streaming a particular video in a low quality constitutes a reduced QoE for a user. However, many VoD streaming services today are used by users for a wide variety of reasons, meaning that user intent cannot be inferred from their use of the service alone. However, our investigation demonstrates that user intent does in most cases coincide with producer intent. We can also demonstrate that the latter can be inferred from the content itself as well as associated metadata. By transitivity, we can choose a default video quality that satisfies the users QoE in the majority of cases.

## Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems:**]: [Video]

## General Terms

Experimentation; measurement; performance

## Keywords

QoE; intent; video streaming

## 1. INTRODUCTION

Video on-demand (VoD) services like Youtube, Vimeo, Netflix, etc. generate most Internet traffic today. It has been predicted that their share will rise to 90% within the next three years[1]. These on-demand videos are used for a wide range of purposes, ranging from entertainment to education but also communication resembling video mail. Currently, VoD streams are delivered at a default quality chosen by the VoD service provider, independent of their purpose. This implies that a user whose intent it is to enjoy exclusively the music of a music video receives the same video quality as a user who wants to enjoy the sights in a nature documentary. There is a discrepancy since delivering a reduced video quality to users with the first intent would not reduce that user's quality of experience (QoE), for the user with the second intent it would reduce QoE. To make this statement, we do not use the term QoE in the spirit of objective video quality metrics, but rather in terms of the International Telecommunication Union's (ITU) formal definition, which defines QoE as "the overall acceptability of an application or service, as perceived subjectively by the end-user" while "the overall acceptability may be influenced by user expectations and context" [1].


**Figure 1: Examples of intent categories. (a) 'Affection': get entertained (e.g. by watching movie); (b)'Experience': learn something (e.g. a recipe); (c)'Information': get informed (e.g. by watching news); (d)'Object': listen to music. Src: Youtube.**

We propose a means by which a VoD service can stream videos with a quality[2] that depends on a user's *context* and the *user expectations* to maximize their QoE. For the study conducted in this paper, we restrict the term *context* to typical knowledge of a VoD service provider, such as the user's age, sex and location. We postulate that these simple criteria are sufficient to identify homogeneous user groups whose intent with respect to use of a particular video are likely to be similar. The classification of users by such criteria is

[1]`http://goo.gl/afWfOH`



*NOSSDAV'15,* Mar 18-20, 2015, Portland, OR, USA

http://dx.doi.org/10.1145/2736084.2736095 .

[2]In this paper, the video quality is expressed in terms of video resolution.

beyond the scope of this paper but they are apparently already exploited by VoD services such as Youtube. Beyond this, *context* includes the situation in which users consume a video stream. Watching news in low quality on a PC monitor in a coffee break may be a satisfactory experience, whereas only a high quality stream satisfies them when watching on a big TV screen at home. The latter challenge has already been explored by analysing user interactions [25].

However, such methods present some limitations: they are mainly designed to determine if the user is interested in both visual and audio content, or the audio content only. If the user is interested in the visual content, the quality that leads to satisfactory QoE may depend on the content itself (e.g. medium for news and high for a movie trailer). User activity may not be sufficient to distinguish these cases.

In this paper, we deal with the specific problem of retrieving the expected quality based on the video content itself. In accordance with our assumptions, we want to establish whether we can deduce QoE from content given the following constraints: (i) *users belong to a single characteristics group*; (ii) *they use the service in the same situation (in their spare time)*; (iii) *they use similar devices (computer with monitor)*. We hypothesize that within these constraints, we can select the lowest satisfactory QoE because we can infer the users' intent, i.e. *why* they watch the video, from the content itself. The proposed solution relies on the three following assumptions: (i) Characteristics of a video such as recording, cutting, encoding, etc., have the potential to reveal the **producer intent** so that it is possible to identify producer intent categories based on the video content; (ii) The producer intent reflects the user intent: the main intent of the person who created and uploaded the video and the one of the person who streams it are similar; (iii)Playback quality that provides satisfactory QoE to the user is directly related the the user's intent.

These assumptions modify the interpretation that has been provided by Hanjalic et al. [10]. While we follow the intent categories that they established, namely 'affection', 'experience', 'information' and 'object', which are explained in Figure 1, we do not postulate that user intent is directly connected to video characteristics. Instead, we postulate that characteristics are expressions of producer intent, and that this provides a good prediction of user intent wherever content is consumed as expected by the producer.

The main contribution of this paper is thus to demonstrate the last two assumptions mentioned above. Firstly, we validate the convergence between producers' intent and users' intent. Secondly, we show that, beyond their ability to classify video content, intent categories reveal the default quality that can satisfies the quality expectations of the user. A proof of concept of the proposed system has been developed to validate our assumptions in a user study.

Last but not least, we demonstrate experimentally that the method has potential to reduce the bandwidth considerably for the delivery of some intent categories, while preserving the user QoE. Although the intent computation is quite error-prone (as our experiments also show), it can be used pragmatically if users are allowed to increase quality manually. In such a scenario, temporary dissatisfaction for some users is tolerated, but considerable bandwidth savings can be achieved compared to the alternative always-best-quality approach, while overall satisfaction is higher than in a hypothetical always-worst-default approach.

In Section 2, we outline works related to QoE considerations in distributed multimedia environments, user intent and resource optimization. In Section 3, a conceptual description of the proposed system (illustrated in Figure 2) is provided. Finally, a validation of the above-mentioned assumptions through a proof of concept implementation of the proposed system is described in Section 4.

## 2. RELATED WORK

Standard internet users are generally not really interested in the technology involved in creating their multimedia content. For most of them, the QoE is the most important concern [12, 11] while watching a video. A lot of research has been done in this direction. For example, Fiedler et al. [9] describe in their work how QoE ties together user perception, experience and expectation to applications and network services. Furthermore, they show how QoE is related to quality of service (QoS).

**QoE considerations.** In the last years, an increase in the number of distributed multimedia environments, devoting particular attention to QoE requirements, has been observed. At the early stage the issue was that, even if they included user involved interaction, the evaluation of these systems was more system centric. Additionally, the proposed approaches were bothersome for the user, due to the fact that users had to provide additional input. Newer concepts tried to change this direction to a more user centric evaluation based on QoE in combination with QoS [24, 20, 20, 21]. This research ranges from providing a general framework to predicting user QoE. Most of the existing research is based on the network layer and the video encoding/decoding process. Krishnan et al. [16] showed that the quality of the video stream can impact the viewers behaviours. In more detail, they showed that rebuffering and startup time of the video can increase the abandonment rate for a given video.

This is an important insight for our work in combination with the fact that people's major concern is video quality (e.g. in terms of video resolution). So, if we can provide users with the content in a quality that satisfies their needs in terms of QoE, it may give the video provider the opportunity to save bandwidth. We try to tackle this problem by connecting the intent of the video producer with the user intent, i.e. why users want to watch the video. We hope that it can both, help providing the user with a better QoE and help allocating bandwidth in a more adapted way.

**User Intent.** User intent has been well investigated in research. In particular research has been done in this direction in textual Web search. [23]. Researchers tried to determine what underlying goal the users have when they use a web search engine [3, 6]. Intent has acquired more and more importance in multimedia research in the last years and multiple studies have tried to make the text retrieval approach usable for multimedia [17, 13, 15]. For example, Lux et al. [19] attempted to find possible intent categories for image retrieval similar to the approach presented in [8]. However, these intent-based papers exploit intent in the context of images. With regards to videos, this issue was treated by Kofler et al. [14], who presented an intent ground truth labelled data set. This is important because they show that, as they exist in the context of image retrieval, user intent categories can be identified in context of video retrieval. Hanjalic et al. [10] write about the intent of videos in the context of video retrieval. They present a cate-

gorization of videos based on the user intent. Further, they provide a method to classify videos based on their intent, and an evaluation of the classification performance.

A newer approach, called intentional framing [22], looks at the framing of images in order to determine the intent of the photographer by analysing the global visual features of the images. The proposed method is strongly related to this approach as we highlight that how a video is produced (e.g. shooted, mounted, etc.) may reflect the producer intent.

**Resource Optimization.** In the context of bandwidth awareness, several methods have been proposed such as means to optimize the ratio between energy consumption and bandwidth. [2, 5]. In Microsoft Azure smooth streaming [25], user behaviour and interaction are utilised to adjust the bandwidth usage, e.g. reducing the quality of the video when the video is in the background or displayed simultaneously with another window. Other researchers looked at the potential of analysing video content in order to adapt the bandwidth usage and the video quality. For example, if there are very complex scenes or a lot of movement in the next frames the capacity needed will be higher [18, 4].

Our work differs from current work in the way that we look at the producers intent in correlation with the quality of the video and the quality of the user experience. To the best of our knowledge, the current state of the art does not provide a solution combining intent and video quality in this way.

# 3. CONCEPTUAL SYSTEM DESCRIPTION

In this section, we describe the general idea and architecture of the system. In order to prove the concept of a multimedia system, able to deliver a content whose quality is related to the producer intent, we implemented parts of the system in a prototype. These parts are described in the experimental section in more detail. The overall system is composed of a client and a server side, shown in Figure 2.

The goal of the proposed system is to understand the user expectations based on the context (if available), the analysis of the video content/metadata and the user behaviour (e.g. via her/his interactions) without requesting any additional information from the user. A typical scenario can be summarized as follows: On the client side, the user is searching for a video while the system is gathering information sent by the user (e.g. text query, url, etc.) on the server side. The playback request then triggers the intent classification. The resulting intent is then considered for deriving a default video quality selection from it. Finally, the video is delivered to the user with respect to the computed quality. If the user is not satisfied with the delivered quality, he/she can change it actively, and any changes in quality settings made by the user is used to feed a semi-supervised machine learning algorithm in order to optimize the expected preferences associated to each intent categories.

**Client Side.** The first characteristic of the client side is that we apply no or little changes to the standard video player functions provided by video platforms. This way, the client side provides the users with a standard interface similar to those commonly used in video services like Youtube and Vimeo. In this interface, the user is in particular able to change the quality of the video (in the same way as it is provided by Youtube). These quality changes, when they occur, are sent to the system and exploited as expected quality "feedbacks", of which the user is unaware. This information can then be used to adjust the quality setting in relation to the intent category. For example, it is possible that, for a certain intent category, the system does not determine the optimal quality setting but, in this case, this determination could be systematically adjusted based on the overall changes performed by the users. The second information that is collected without awareness of the user is the behaviour of the user regarding the focus of the windows. An example is the video presented in a window in the background and not actively shown, which is strongly connected to the approach from Azure Smooth streaming [25]. This information is particularly useful in detecting certain types of intent categories, e.g., listen to music or a podcast. Information that is not unconsciously provided but a natural input of the user in a video search engine is the search query. Finally, other user-related information can be collected from the user side such as available bandwidth, used display device, etc.

**Server Side.** On the server side, we implemented so far the intent classification and the intent-to-quality mapping. All other parts are described on a conceptual level. The main system consists of four parts. The first part is the producer intent classification which is responsible for placing each of the videos into one of the intent categories. This is done based on the method presented in [10]. This approach is described in terms of user intent, we adopt it here in order to detect the producer intent. To classify the intent, different sources of information have to be analysed, i.e. the visual and audio content, the metadata and the user input and feedback. The results of this classification will then be combined by late fusion. The second part is the video search engine. This is important because the search query itself can be a valuable source of information about the intent. The third part is the quality association part. This part uses the intent information from the intent classifier to determine the quality of the video delivered to the users. It manages the used codecs but also the final resolution. It creates an intent-quality model that tries to determine the quality of the videos based on the constraint that the bandwidth allocation must be optimized. Furthermore, it also learns from the users feedback (based on whether or not the quality is changed). The last part is the bandwidth thresholding part. This part is responsible for the optimization of the bandwidth usage and is based on the intent and available bandwidth information. It is important to point out that our system will not try to give the user the best quality based on the bandwidth available. It is more a new way to look at the distribution and usage of bandwidth by trying to satisfy the user needs based on the intent without wasting bandwidth.

# 4. EXPERIMENT

The idea of this initial experiment is to show the convergence between producer intent and user intent, and that different intent categories are correlated to different video quality expectations and therefore bandwidth allocation.

The experiment is split into two parts. The first part is the automatic clustering based on several features of the videos. Since our system is not complete, these initial experiments will show whether building such a system is sensible. The clustering and feature extraction are based on well known methods and frameworks. Moreover, we want to point out that, in future work, we will develop and implement more

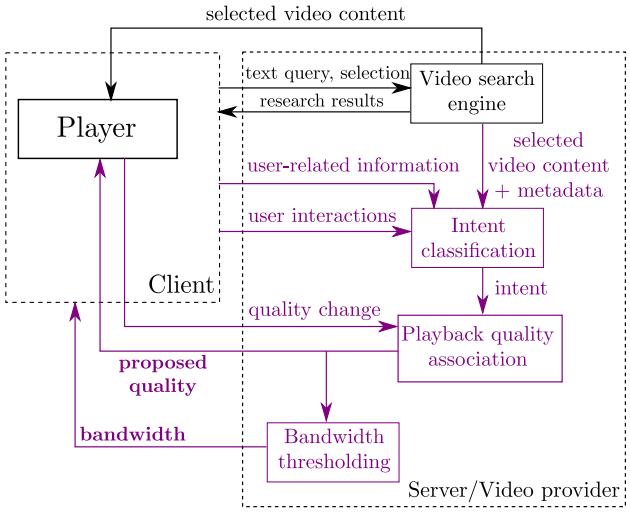



**Figure 2: Overview of the proposed system composed of a client side (left) and a server side (right).**

sophisticated methods. The second part of the experiment is the user test where we show that the producer intent is correlated to the user intent and that intent is somehow related to the video quality that satisfies the user QoE.

The intent classification part is based on [10]. The classes for the classification are 'Information', 'Experience', 'Affection' and 'Object',i.e. in our context, listen to music. The automatic clustering is performed based on audio and visual features and metadata, which consist of title, description and tags. For the audio information, we used ASR (Automatic Speech Recognition) and for the visual features we used Shot Patterns algorithm. For clustering, we used the Weka machine learning framework [3] and the K-means clustering algorithm. We first calculate the possible cluster for each feature and then combine them in a late fusion step.

For the second part we developed an HTML video player which allowed us to control the quality setting of the video and in the same time collect feedback from the users. We then used a set of 10 trusted users (who we knew would perform the task accurately). Because of the low amount of participants we decided for a placebo-controlled study to make it more to robust. Therefore, we split these 10 users into two equal groups. One group, referred to as by *real group* in the rest of the paper, got videos with quality settings based on the producers intent. The other 5 users, referred to as by *placebo group*, got the same videos with the default level of quality (that we defined as medium with 360p). We chose this way to implement our experiment, in order to, compare the two groups and asses if our method successfully improved the QoE and bandwidth usage combination compared to the standard settings, thus making the experiment more robust.

We downloaded a set of 400 random videos from Youtube that we clustered into the 4 different intent categories. We modified the description of the intents in a way that they are easier to understand for the user. In our case, we chose "listen to music" for the 'Object' intent category. One will maybe assume that music is related to entertainment; this is partially true but music can not be reduced to just entertainment as many people use music for other purposes such as *get relaxed* or *support them at work*. After the classification of the videos we randomly chose 5 videos per intent class. This leaded to a dataset of 20 videos in total for the user test. They range from *cinema trailers* to videos about *how to learn Japanese*. Most of them have a clear intent category. Some can be in more than one category in which case we asked the users for the most fitting one. The video duration varied from some minutes to almost one hour. For the quality representation, we used the Youtube standard settings which are small (240p), medium (360p), large (480p) and hd720 (720p). We did not use higher resolution than 720p because not all of the videos supported it.

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

**Table 1: This table shows the users opinion about the producer intent of the videos in the experiment.**

| Users Intent / Classified intent | Affection | Experience | Information | Object: listen to music |
|---|---|---|---|---|
| Affection | 46 | 1 | 0 | 3 |
| Experience | 2 | 44 | 2 | 2 |
| Information | 2 | 8 | 38 | 2 |
| Object: listen to music | 11 | 1 | 0 | 38 |

**Table 2: This table depicts the users satisfaction and used bandwidth in MB. Each column presents one intent category (affection, experience, information, object).**

| Real group | | | | Placebo group | | | | Used Bandwidth in MB | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Preset quality | yes | higher | lower | Preset quality | yes | higher | lower | small | medium | large | hd720 |
| hd720 | 24 | 0 | 1 | medium | 1 | **24** | 0 | 49.31 | 93.4 | 141.1 | 318 |
| large | 11 | 1 | **13** | medium | **18** | 4 | 3 | 22.29 | 64.98 | 93.37 | 265.6 |
| medium | **14** | 1 | **10** | medium | **12** | 6 | **7** | 22.1 | 57.2 | 80.6 | 201.8 |
| small | **21** | 3 | 1 | medium | 0 | 7 | **18** | 13.26 | 26.93 | 38.16 | 81.4 |

We then randomly assigned the videos to either the placebo or the real group and each user had to watch all 20 videos and indicate which intent they would choose for each video as well as whether or not they were satisfied with the video quality they were provided with. In introduction to the experiment, a clear and user friendly description of the four intent categories was outlined. In order to insure the correct execution of the assignment, the clarity of the formulation was assessed by performing preliminary tests with five different users. Since we wanted the users to consider the video quality in detail, we formulated the question in a way that arrogates this behaviour. The question for the quality was: *Are you satisfied with the visual quality of the video?*. The possible answers were (i) I would like to watch the video in a higher quality, (ii) I would watch the video in lower quality and (iii) Neither 1 nor 2.

## 4.1 Results

The collected information support our assumption that producer intent is related to user intent. In consequence, this result suggests that it may be possible to exploit the relation between the producer intent for a video and the quality expected by the user. An overview of the results can be found in Table 1 and 2. The first table contains the opinion of the users from both groups about the producer intent of a video. It can be seen that the users' opinion agrees in majority with the producers intent for the video. The second table shows the user opinions about the quality for each test group and summarizes bandwidth usage in MB per intent class and quality levels for all videos.

**Affection.** For the affection intent class and in both groups (real and placebo), the participants agreed clearly on the producer intent question. In the group that got the quality settings based on our system the users were satisfied with the quality. They only voted with *yes we are satisfied* or *higher quality*, which we count as satisfied because we set the maximum available quality for the video. In the placebo group, only one user was satisfied with the quality. All other users wanted to watch the video in higher quality, which shows that, the medium quality setting does not satisfy the users quality of experience needs for this intent. In this case, the system uses more bandwidth (compare the last four columns of 2) but the users satisfaction is higher compared to the placebo group.

**Experience.** For the experience class, we got completely different results as expected. We set the quality for this videos too high (because we assumed that when one experiences something they may want to do it in good quality). For both groups, the opinion about the intent of the videos was clear. The majority of the users in the real test group voted for lower quality. In the placebo group, they were always satisfied with the medium quality (which is one step lower than in the real test group). This gave us two interesting insights. First, the intent of experience is not related with the large quality setting requirements. Secondly, taking the users feedback into account will help improving our system in the future.

Another interesting point was that, one of the videos was an outlier in both groups (affection instead of experience). In the real group, the users were satisfied with the large quality or they expressed the wish for a higher quality. In the placebo group, a higher quality than medium would have been preferred for this particular video. The video was about someone who was playing a computer game and recorded it. This type of videos, called *lets play*, are becoming more and more popular in recent years [4] and are made by the producers for entertainment and not learning purposes. There also exist video platforms which specialize in this type of video [5]. It could definitely also be a video that teaches how to play the game, but such video would have different features regarding content and user-related information.

**Information.** For the information intent category, we had in both groups a high satisfaction rate. This is justified by the fact that, the medium quality setting was chosen for this intent, which also corresponds to the default setting of the placebo group. Moreover, we had a high precision for the producer vs. user intent classification. An outlier, which was a video about learning Japanese, has been misclassified by our system. This video should be in the intent category of experience/learn something. Another important observation was that, it seems that, user would be satisfied with even a lower resolution than medium for the information intent category. The bandwidth saving potential of this intent category could be even higher.

**Object: Listen to Music.** The experiment showed us that for this intent category, the lowest playback quality provides satisfying QoE. This can consequently be a very efficient way to save bandwidth without reducing the QoE for the users. An outlier was observed. It was a scene from the *Lord of the Rings* movies, where a Hobbit is singing a song to the lord of a city. Almost all participants voted for this video *affection* as their intent, and they also wanted to see it in a higher resolution, even if most of the part of the clip is a song sung by the Hobbit. We consider this as an indicator that at first, producers intent is very hard to detect. And second, that we definitely need information about the context to be more accurate in the classification part of the system. Finally, the last column in Table 2 reveals, for this intent category, a high potential for saving bandwidth while preserving a playback quality that satisfies the users QoE.

# 5. DISCUSSION

The experimental results emphasize the convergence between producer intent and user intent. Furthermore, they also agree with the hypothesis that intent information can be exploited in order to adapt the bandwidth distribution. The user votes on the quality satisfaction gave indeed an indication that intent categories are related to quality expectations. It also showed that these categories can help to satisfy the users QoE, either by simply improving it, e.g. with higher video quality, or by preserving it while decreasing the default playback quality. Furthermore, we showed that exploiting these intent information can be a promising idea for interesting bandwidth allocation and saving as it can be seen in Table 2. This can be done based on the fact that videos can be assigned to different intent categories. An allocation based on these intents could help to share bandwidth in a way that the QoE is maximized over all users, or at least, group of users. This would prevent to waste bandwidth by always trying to provide the highest possible quality. This could lead to another important side effect namely saving energy.

The problem of saving energy in the context of online videos has been recently addressed in [7]. The authors looked at the energy consumption caused by different video codecs and video resolutions. A potential problem regarding the applicability of the proposed method is that they have to increase end user awareness and somehow interact with her/him. This can be a challenge as users are generally unwilling when it comes to providing additional information which are not directly associated with their initial goal (streaming a video). Since our system can work independently to the user willingness to cooperate, it could be interesting to further exploit our approach, now for its energy saving potential.

Furthermore, in terms of implementation, it could be very interesting to use DASH [6]. In this scenario, the angle would be changed from just how-much-bandwidth-do-we-have based methods to something more user centred. For example, lets assume the system knows that it has 60 users which want to get entertained, but also 200 users who just want to listen to music. A higher bandwidth (for the better quality) would then only be needed for 60 users who have a need for it, and the rest could be satisfied with the lower bandwidth. This information could be used at the point when the system allocates the bandwidth for the users. It would be an especially interesting alternative when we take the global aspect and the billion of possible users into account and thus its great bandwidth and energy saving potential. This paper is of course just a small step in this direction but the most important insight is the relation between the quality expected

[4] `http://goo.gl/YrvnWf`
[5] `http://www.twitch.tv/`
[6] `http://dashif.org/mpeg-dash/`

by the user and the producer intent, which reflects the user intent in most cases, as a valuable source of information.

A possible limitation of the proposed method is the fact that, it only makes sense in services where the user can search for videos freely like Youtube or Vimeo. Using it in services like Netflix or HBO which have a clear intent before the users start using the service, i.e. in that case *get entertained*, does not seem useful. However, it can not be seen as completely useless because the insights of such a system may be used by these very specialized portals to improve the QoE of their users. For example, it may be interesting to systematically provide low level quality videos on a news video service because the majority of the users will be satisfied with the lower quality level.

Finally, we want to point out that our approach is not only based on the user behaviour or the content. As observed in the experiment section, the user tests showed that users accept lower quality for videos with the intent of information or experience. For these videos, the *being in the background*, *just partially visible* or *just looking at the content* approach would not work well or at all, because it misses the real understanding of the user need. In that case, it makes sense to look at the producer intent. The two approaches are of course complementary and the idea is to adapt the system based on the users feedback but mainly in the sense of learning the intent and the lowest video quality acceptable without impiding a good QoE. Therefore, in a way, we secretly entice the user to using a lower quality without letting them be aware of it. Of course, there will be users that will be unsatisfied and increase the quality but if the main part of the users accept it, bandwidth will still be saved.

# 6. CONCLUSION

We presented a novel system able to detect a social signal, namely the producers intent and showed that it is related to the users intent for watching a video. We discussed it in context of potential bandwidth and energy saving. The detection of the intent is based on the content, metadata and user related information. Based on the partially implemented system classification, we provided different quality levels to the user. We performed a user study that revealed, that users agree about the producers intent and that they were more satisfied by our system preset qualities than the standard quality setting. This is a strong indicator that such a system can be a new way to look at means to provide content to the users. The next steps include collecting a large scale dataset and conduct experiments over a longer period of time. In future experiments we will also collect information about bandwidth and energy usage levels. This will give us more accurate insight of the possible saving potential.

# 7. ACKNOWLEDGEMENTS

This work is partly funded by the FRINATEK project "EONS" (#231687) and the iAD Centre for Research-based Innovation (#174867) by the Norwegian Research Council. We also want to thank Vamsidhar R. Gaddam for discussion.

# 8. REFERENCES

[1] Vocabulary for performance and quality of service. In *ITU-T Rec.incl. Amendment 2*, 2008.

[2] C. Bae and W. Stark. Energy and bandwidth efficiency in wireless networks. In *Proc. of CCS*, volume 2, pages 1297–1302. IEEE, 2006.

[3] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. The intention behind web queries. In *String processing and information retrieval*, pages 98–109. Springer, 2006.

[4] A. Begen, T. Akgul, and M. Baugher. Watching video over the web: Part 2: Applications, standardization, and open issues. *IC, IEEE*, 15(3):59–63, 2011.

[5] K. Bhardwaj and R. K. Jena. Energy and bandwidth aware mapping of ips onto regular noc architectures using multi-objective genetic algorithms. In *Proc. of SOC'09*, pages 027–031. IEEE, 2009.

[6] A. Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[7] O. Ejembi and S. N. Bhatti. Help save the planet: Please do adjust your picture. In *Proc. of the ACM MM*, pages 427–436. ACM, 2014.

[8] R. Fidel. The image retrieval task: implications for the design and evaluation of image databases. *NRHM*, 3(1):181–199, 1997.

[9] M. Fiedler, T. Hossfeld, and P. Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. *IEEE NW*, 24(2):36–41, 2010.

[10] A. Hanjalic, C. Kofler, and M. Larson. Intent and its discontents: the user at the wheel of the online video search engine. In *Proc. of the ACM MM*, pages 1239–1248. ACM, 2012.

[11] G. Harman. The intrinsic quality of experience. *Phil. persp.*, pages 31–52, 1990.

[12] R. Jain. Quality of experience. *IEEE MM*, pages 96–95, 2004.

[13] B. J. Jansen, A. Spink, and J. O. Pedersen. The effect of specialized multimedia collections on web searching. *Web Eng.*, 3(3-4):182–199, 2004.

[14] C. Kofler, M. Larson, and A. Hanjalic. First version of an intent ground-truth labeled data set.

[15] C. Kofler and M. Lux. Dynamic presentation adaptation based on user intent classification. In *Proc. of the ACM MM*, pages 1117–1118. ACM, 2009.

[16] S. S. Krishnan and R. K. Sitaraman. Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs. *IEEE/ACM TN*, 21(6):2001–2014, Dec. 2013.

[17] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM TOMCCAP*, 2(1):1–19, 2006.

[18] Z. Li, A. C. Begen, J. Gahm, Y. Shan, B. Osler, and D. Oran. Streaming video over http with consistent quality. In *Proc. of ACM MMSys*, pages 248–258. ACM, 2014.

[19] M. Lux, C. Kofler, and O. Marques. A classification scheme for user intentions in image search. In *CHI'10*, pages 3913–3918. ACM, 2010.

[20] V. Menkovski, A. Oredope, A. Liotta, and A. C. Sánchez. Predicting quality of experience in multimedia streaming. In *Proc. of the AMCM*, pages 52–59. ACM, 2009.

[21] S. Moller, K.-P. Engelbrecht, C. Kuhnel, I. Wechsung, and B. Weiss. A taxonomy of quality of service and quality of experience of multimodal human-machine interaction. In *Proc. of the QoMEx*, pages 7–12. IEEE, 2009.

[22] 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 the ACM MM*, MM '14, pages 397–406, New York, NY, USA, 2014. ACM.

[23] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proc. of WWW*, pages 13–19. ACM, 2004.

[24] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang. Quality of experience in distributed interactive multimedia environments: Toward a theoretical framework. In *Proc. of the ACM MM*, MM '09, pages 481–490, New York, NY, USA, 2009. ACM.

[25] A. Zambelli. Iis smooth streaming technical overview. *MS Corp.*, 3, 2009.

# Paper IV

# Media Synchronization and Sub-Event Detection in Multi-User Image Collections

# Media Synchronization and Sub-Event Detection in Multi-User Image Collections

Maia Zaharieva[1,2] Michael Riegler[3]

[1]University of Vienna, Austria
[2]Vienna University of Technology, Austria
[3]Simula Research Laboratory AS and University of Oslo, Norway
maia.zaharieva@univie.ac.at, michael@simula.no

## ABSTRACT

Personal media capturing devices, such as smartphones or personal image and video cameras, are rarely synchronized. As a result, common tasks, like event detection and summarization across different multi-user media galleries, are considerably impeded and error-prone. In this paper, we investigate different approaches for the synchronization of image collections using visual information only. We perform a thorough evaluation of the performance of several global features on three datasets. Additionally, we explore the feasibility of common clustering algorithms for the detection of sub-events in the presence of synchronization misalignment.

## Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Search and Retrieval; H.3 [**Information Systems**]: Content Analysis and Indexing; I.4 [**Image Processing and Computer Vision**]: Applications

## General Terms

Experimentation

## Keywords

Media synchronization, sub-event detection, multi-user image collection, MediaEval

## 1. INTRODUCTION

Nowadays, a large amount of people own a device that can take pictures or videos. Such devices range from professional cameras to smartphones. In a lot of events, such as sport events and music festivals, people produce and share media capturing these events. Moreover, the images and videos taken from the same event but from different users do not just capture the same event but also different angles and details. Such data hold a lot of potential. Imagine, for example, a birthday party where several people take pictures and videos. At the end of the day the person who had her birthday could get a perfect summarization of the event based on this data. This scenario is not just limited to social events. The same approach could be used for concerts, sport events, and any other type of event. Beside all the advantages this data provides, it also bears some challenges ranging from mining large media collection to diversification and summarization of heterogeneous data. A crucial issue in this context is the synchronization of multi-user data in terms of time. Different devices have different clocks and time zone settings or even potential malfunctions. In general, it is not feasible, that users participating in the same event synchronize their devices in advance. This leads to collections of multimedia data for the same event that can differ considerably from minutes or hours to even days. As a result, if the timestamps of all this data is not sychronized, further tasks like event detection, summarization, and clustering, get error-prone.

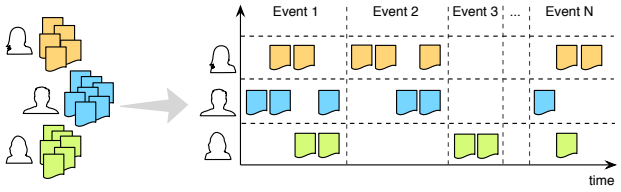
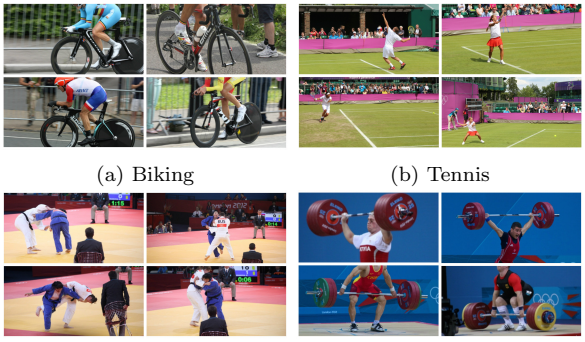



Figure 1: Application scenario.

In this paper, we propose two simple but efficient approaches based on contextual information and global visual features that facilitate the sychronization problem in the context of multi-user image collections. Additionally, we present approaches to cluster a big event, in our case Olympic games, in several sub-events. We perform extended experiments to assess the potentials and boundaries of the employed approaches in a combination with a broad range of well-established visual features. Figure 1 presents a general visualization of the application scenario. We assume, a group of people attending and capturing the same event, whereas not all users need to take part in the same sub-events. We explore two different approaches for the synchronization of images across the different, user-based collections. In the following, we investigate the feasibility of visual-based clustering algorithms. Figure 2 shows, as an example, different sub-events detected by one of the employed sub-event detection methods.


*HuEvent'15,* October 30 2015, Brisbane, Australia.

978-1-4503-3748-9/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2815244.2815248.



(a) Biking (b) Tennis

(c) Judo (d) Weightlifting

Figure 2: Example results for detected sub-events detected.

This paper is organized as follows. Section 2 provides a brief overview of related work in the context of multi-user media synchronization and event detection. Section 3 describes the proposed approaches in detail. Section 4 presents the performed experiments and the obtained results. We conclude the paper in Section 5.

# 2. RELATED WORK

## 2.1 Synchronsiation of Multimedia Events

Synchronization of multimedia content has been discussed in literature for a long time [3]. However, recent research shows, that the problem itself is far away from being solved due to new types of media and technologies challenging the existing approaches. For example, Broile et al. employed the image content to estimate mutual delays between different cameras [5]. The authors used region- and color-based matching in a combination with a local feature descriptor (Speeded Up Robust Features, SURF[2]) for salient point matching and delay estimation. They report experimental results based on a user-generated dataset. The proposed algorithm synchronized about 80% of the considered galleries within a delay of two minutes. However, the use of local image features can be very time consuming when it comes to large collections. Recently, Veenhuizen and Brandenburg explored frame accurate media synchronization [19]. The authors employ a collection of pipelines, each of which provides a common clock to its elements. In a combination with the presentation timestamp for each frame,the pipelines are used for precise synchronization. To evaluate their approach, the authors took a video and cut it in a left and right part. This makes it easy to manually check for synchronization errors. The authors recommend the use of time information; however, this information is not available in the context of image collections. Yang et al. used a kernel space graph, visual features (color histogram, Scale Invariant Feature Transform (SIFT)[7], GIST[11], locality-constrained linear coding (LLC)[20]), and location information where available to sychronize image galleries of different users for the same event [21]. The results are evaluated on a self-collected dataset with 36 image galleries. The evaluation part does not provide many results; however, the authors state that they can reach the expectation of consumers without providing any further details. Recently, Sansone et al. presented an approach for the sychronization problem using a probabilistic graphical model [18]. Based on a set of nearest neighbour pictures, a temporal displacement is identified. This information is used to calculate the offset by means of a graphical model using global color structure descriptors (CSD)[1] and local binary patterns (LBP)[10].

## 2.2 Event Detection in Multimedia

Event detection in multimedia content, such as image galleries, streams, and video collections, is a subject to very active research recently. For example, Petkos et al. use multimodal spectral clustering and early fusion to combine several heterogeneous features [15]. By using an explicit supervisory signal, good clustering accuracy is reported. Reuter and Cimiano classify social media streams into corresponding events [16]. The authors specifically focus on the detection of new events and on the assignment of detected events to an already known event. Furthermore, the authors address the scaling of the data and report high quality and scalability of the proposed approach. Recently, Papagiannopoulou and Mezaris aim at the summarization of image collections into single events [13]. The authors employ a clustering approach based on concept detection. The combination of clustering and concept detection leads to a clustering of events that is quite similar to human expectations. Recently, Sansone et al. used their algorithm for clustering on the same datasets employed in this paper [18]. The authors report F1-scores of 0.14 and 0.11 for the Vancouver and London2 datasets respectively, which demonstrates the challenges of the employed data. Additionally, the authors employ k-means clustering, which requires for prior knowledge or an estimation of the number of events.

# 3. APPROACH

## 3.1 Synchronization of Image Galleries

We investigate two approaches for the synchronization of image galleries with respect to the number of reference images to estimate a time offset. While the first approach builds upon pairwise distances (PTS, pairwise transitive synchronization), the second approach considers groups of images (CBS, cluster-based synchronization).

### *3.1.1 PTS: Pairwise Transitive Synchronization*

Our first approach for the synchronization of image galleries considers pairs of visually highly similar images to construct a transitive list of entry points to all galleries (including the reference gallery). For this purpose, we first construct the pairwise similarities between all images of all galleries. We sort the pairs of images in ascending order according to their dissimilarity level. We process this list and consider the images of a pair identical if 1) they originate from different galleries and 2) their similarity level is within a predefined threshold. Images, representing different galleries are declared as entry points for the synchronization of the corresponding galleries. We proceed with the sorted pair list until we are able to build a transitive list of entry points to all galleries represented in the full dataset or we reach the end of the list. Eventually, all galleries are timely aligned according to the provided reference collection using the corresponding entry points.

### *3.1.2 CBS: Cluster-Based Synchronization*

The second approach considers groups of visually similar images to estimate the time offset between galleries. First, we cluster all images using an unsupervised clustering method which does not require prior knowledge of the

number of underlying clusters (e.g., X-Means [14]). For each cluster, we consider the average deviation of the timestamps of the reference images to all other images of a collection represented in the same cluster as the offset of this image collection. In the case, that there are less than two reference images in a cluster, we use the available corrected timestamp of non-reference images which already have an offset from another cluster. For inconsistent time information from different clusters, we use the cluster with highest number of time information (similar to a simple majority vote).

## 3.2 Sub-Event Detection in Image Galleries

In our application scenario, user-generated image galleries for primarily personal use, the available information is commonly limited to the information generated by the camera, i.e. capture time and potentially GPS data, and the visual content itself. Therefore, we first perform sub-event detection in the image galleries using clustering based on the visual information only. Additionally, we explore the combination of visual information and time or GPS information, i.e. two visual clusters are merged if 1) they share a common gallery and 2) the minimum time/GPS distance between them is below a predefined threshold. This step is repeated until no further cluster merging is possible. This process assures that clusters can grow successively and that detected sub-events are not limited to a predefined time duration or spatial location.

# 4. EXPERIMENTS

In this section we present the results of the performed evaluation on the synchronization of multi-user image galleries and on the detection of potential sub-events within the provided galleries. Results on the synchronization of the image collections are reported in terms of average deviation in seconds between the real and the estimated time offset of a gallery with respect to the reference gallery. Results on sub-event detection are reported in terms of recall ($R$), precision ($P$), and F1-score ($F1$) representing the harmonic mean between recall and precision, and normalized mutual information ($NMI$) measuring the goodness of the performed clustering with respect to the ground truth.

## 4.1 Dataset

For our experiments, we employ a dataset that was successfully used as part of the MediaEval Benchmark 2014 task on the synchronization of multi-user event media (SEM) [6]. The data is arranged in three sets: two sets consisting of Flickr images from the *London Olympic Games 2012* and one set with images from the *Winter Olympic Games in Vancouver, 2010*. For each data set, one gallery is declared as the reference gallery and all other galleries have to be synchronized with respect to it. Each gallery within the dataset is consistent in terms of timestamp since it is captured by the same user. However, the time offset across different galleries is partly considerable.

Table 1 summarizes the data characteristics within the different sets. Additionally to the strongly varying time offset, the employed datasets differ significantly in the number of included galleries and in the number and distribution of images. Eventually, approximately 50% of all images in the two London datasets and more than 67% of the Vancouver set provide location information. The underlying sub-events in the datasets in terms of number of events and image distribution show notable variations. While the London1 dataset is the smallest one with 304 images distributed over 10 galleries, it covers a relatively large number of sub-events (58) with partly few number of images. On the opposite, the Vancouver dataset covers 86 sub-events with a higher number of images. Please note, that in the context of this work we employ the term sub-events, because an event is a broader term that would commonly refer to the Olympic Games as a whole. Sub-events in the investigated datasets describe different parts of the major event, such as *opening section*, *national anthem*, *parade of the nations*, etc. While the Olympic Games last for several days, the identified sub-events are highly granular and of comparatively very short duration, e.g. a few minutes.

Table 1: Dataset overview (G: number of galleries, E: number of events, $\mu$: mean, $\sigma$: standard deviation).

| Dataset | G | Images | | | Offset in sec. | | GPS | E | Images/Event | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | total | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | (%) | | $\mu$ | $\sigma$ |
| London1 | 10 | 304 | 30.40 | 5.19 | -3,000 | 12,914 | 52.63 | 58 | 5.24 | 5.30 |
| London2 | 37 | 2,124 | 57.41 | 77.06 | 3,003 | 17,600 | 51.04 | 238 | 8.92 | 14.56 |
| Vancouver | 35 | 1,351 | 38.60 | 13.03 | 1,242 | 15,570 | 67.23 | 86 | 15.53 | 15.91 |

Table 2: Overview of employed features (in alphabetical order) and the corresponding dimensionality (D).

| | Feature | D |
|---|---|---|
| ACC | Auto Color Correlogram | 1,024 |
| CEDD | Color and Edge Directivity Descriptor | 144 |
| CL | Color Layout | 192 |
| FCTH | Fuzzy Color and Texture Histogram | 192 |
| JCD | Joint Composite Descriptor | 168 |
| JCH | Jpeg Coefficient Histogram | 192 |
| M7CL | MPEG-7 Color Layout | 27 |
| M7CS | MPEG-7 Color Structure | 256 |
| M7EH | MPEG-7 Edge Histogram | 80 |
| M7HT | MPEG-7 Homogeneous Texture | 62 |
| M7RS | MPEG-7 Region-based Shape | 35 |

## 4.2 Visual Features

In this evaluation, we focus on global visual features that have been widely used for image similarity estimation. The employed features represent different visual aspects, such as color, structure, and shape information. We selected these features, because they are lightweight in terms of computational cost and size (dimensionality) which makes them very suitable for tasks with a lot of data. Additionally, recent research showed that they can perform at least as good as local visual features like for example SIFT or SURF [17]. Table 2 provides an overview of the features we consider for comparison in our experiments. For background information on the features, please refer to [4][8][9].

## 4.3 Synchronization of Image Galleries

In our first experiment, we compare the performance of the two proposed approaches for media synchronization, PTS and CBS, using different visual features as presented in the previous section. For the CBS approach, we employ the X-Means clustering algorithm [14]; however, any other clustering method can be employed as well. Table 3 summarizes the results for the three employed datasets in terms of average deviation in seconds between the real (ground truth), $o_r$, and the estimated, $o_e$, time offset. In order to reflect different aspects of the results, we report both relative and absolute deviations. While the average relative deviation is more precise, $\mu_r = 1/k \sum_{i=1}^{k} (o_r - o_e)$, where

$k$ is the number of galleries in the corresponding dataset, the absolute deviation provides an overall estimation of the performance independently of the sign of the underlying deviation, $\mu_r = 1/k \sum_{i=1}^{k} |o_r - o_e|$. As a baseline, we consider the provided offsets in the datasets, i.e. what would be the result if no synchronization at all is performed.

The results show that there is no single best solution for all three datasets. While the PTS approach achieves a significant correction of the time offset on *London1* and *Vancouver* (e.g. 21 seconds vs. 3,000 seconds on London1), the performance on *London2* remains far behind the baseline. On the opposite, CBS seems to handle *London2* data better than the PTS approach. The core reason for the partly substantial difference in the performance on *London2* is the nature of the underlying data. This dataset contains visually highly similar sub-events taking part on different days (e.g. marathon vs. 10,000 meters running). As a result, a single incorrect assignment may easily propagate. In order to account for such cases, the temporal consistency of the assignments needs to be additionally considered. The results show that there is no distinct tendency towards a given feature. Nevertheless, the ACC and M7CS features are among the top performing features for all datasets, while FCTH, M7EH, M7HT, and M7RS show, in general, lower performance. Overall, the PTS approach outperforms CBS in terms of both relative and absolute deviation. The difference becomes more obvious when the distribution of the time deviations is investigated in more detail (see Figure 3). The synchronization of the most galleries using the CBS approach results in a deviation from the ground truth from more than an hour. On the opposite, PTS demonstrates more precise and reliable synchronization using the top features with the majority of the galleries synchronized within ten minutes or less from the ground truth.

## 4.4 Sub-Event Detection in Image Galleries

In our second experiment, we explore time and location information as well as the visual content of the images in order to identify sub-events in the investigated datasets.

We first investigate the performance of purely visual-based clustering. We consider two unsupervised clustering approaches, Agglomerative Hierarchical Clustering (AHC) and X-Means, which automatically estimate the number of clusters. The results show, that the AHC settings (single linkage, cutoff parameter of 1) are very restrictive resulting in a strong over-segmentation as the most clusters consist of only two images (c.p. the number of clusters in Table 4). However, the high precision indicates that detected clusters tend to be pure. On the opposite, X-Means detects much lower number of clusters and often leads to under-segmentation. This is due to the partly high visual similarity among different sub-events (e.g., same discipline but different teams or different classes). As a result, both the precision and the overall F1-score remain mostly considerably low. This indicates that a more strict visual-based clustering will potentially provide a more reliable foundation for further analysis. Additionally, the results show that there is no single best performing visual feature: the top three features in terms of averaged ranking for the three datasets are ACC, JCD, and M7CS for the AHC clustering method and CL, JCD, and M7CS for the X-Means clustering. Due to space limitations, for the combination of visual and time-based clustering, we focus only on the common set of top performing features: ACC, CL, JCD, and M7CS.

In a next step, we explore the combination of time- and visual-based information after considering the correction of the time offset from the event synchronization step. The simplest feature combination is the early fusion of the two feature types. The results show, that independently of the clustering approach, the performance is comparable to those of the purely visual-based event detection. Since the AHC clustering results in high over-segmentation, we additionally employed the time-based refinement step as presented in Section 3.2. The approach merges visual-based clusters if they share images of the same user (same gallery) and if the corresponding time distance is below a predefined threshold. In such a way, detected events can be of varying temporal duration and no assumption about the potential event duration in the employed dataset has to be made. We explored different settings for the time threshold ranging from 1 to 120 minute. Due to space limitations, we only present the top results achieved for a time threshold of 3 minutes. Overall, both the number of detected clusters is considerably reduced and the performance in terms of F1-score is improved for all employed features. However, the precision is partly notably reduced as a result of falsely merged event clusters. This indicates that the underlying sub-events are not well-separated but timely consecutive or even potentially overlapping.

Eventually, we explore the location information, where available, as an alternative refinement step for the visual-based clustering. Again, we employed different thresholds for the location-based distance ranging from 10 to 500 meters. Overall, best performance was achieved for a threshold of 30 meters. Again, the performance improves considerably in comparison to the purely visual-based clustering both in terms of F1-score and number of detected event clusters. However, the slightly reduced precision indicates, that sub-events are also spatially not well-separated. This is mainly due to the high granularity of definition of sub-events, e.g. different parts of the opening ceremony.

# 5. CONCLUSION

In this paper we addressed two fundamental tasks in the context of mining multi-user image collections using available contextual and visual information: media synchronization and sub-event detection. The proposed approach in context of media synchronization achieves an outstanding performance for two out of the three investigated datasets. However, the thorough investigation of sub-event detection indicates the boundaries of visual- and contextual-based clustering in this context. Further consideration of potentially available textual descriptions or external sources of information is required in order to further improve the results. Additionally, we will investigate the impact of different clustering algorithms since recent studies suggest, that graph-based clustering leads to higher-quality clusters in comparison to conventional clustering methods [12].

# Acknowledgments

This work has been partly funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-010 and by the Norwegian FRINATEK project "EONS" (231687).

# References

[1] M. Baştan, H. Çam, U. Güdükbay, and O. Ulusoy. Bilvideo-7: an MPEG-7-compatible video indexing and retrieval system. *IEEE MultiMedia*, 17(3):62–73, 2010.

Table 3: Synchronization results in terms of average deviation in seconds between the real and the estimated time offset ($\mu$) and the corresponding standard deviation ($\sigma$). Bold values indicate best performance in terms of $\mu$ for each dataset and each synchronization approach.

| Approach | Feature | Relative Deviation | | | | | | Absolute Deviation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | London1 | | London2 | | Vancouver | | London1 | | London2 | | Vancouver | |
| | | $\mu_r$ | $\sigma_r$ | $\mu_r$ | $\sigma_r$ | $\mu_r$ | $\sigma_r$ | $\mu_a$ | $\sigma_a$ | $\mu_a$ | $\sigma_a$ | $\mu_a$ | $\sigma_a$ |
| Baseline | – | -3,000 | 12,914 | 3,003 | 17,600 | 1,242 | 15,570 | 9,400 | 8,803 | 14,037 | 10,788 | 12,162 | 9,571 |
| PTS | **ACC** | -220 | 259 | 34,729 | 166,720 | -2,011 | 3,687 | **220** | 259 | **34,606** | 161,700 | 2,877 | 3,039 |
| | CEDD | -1,338 | 1,311 | -164,966 | 480,333 | -39.014 | 108,864 | 1,338 | 1,311, | 301,261 | 496,645 | 44,485 | 106,681 |
| | CL | 81 | 942 | -99,816 | 293,418 | -37,786 | 183,573 | 748 | 515 | 112,185 | 288,782 | 68,473 | 174,182 |
| | FCTH | -262,303 | 313,723 | -37,387 | 144,494 | -19,805 | 121,903 | 264,803 | 311,352 | 43,736 | 142,648 | 22,099 | 121,496 |
| | JCD | 457 | 1,277 | -387,873 | 582,364 | 2,977 | 4,388 | 1,124 | 666 | 430,428 | 537,904 | 3,295 | 4,147 |
| | JCH | -383 | 1,209 | -91,726 | 361,194 | 25,195 | 61,712 | 498 | 1,160 | 168,977 | 331,280 | 25,943 | 61,391 |
| | M7CL | -299 | 1,002 | -60,132 | 284,149 | -32,014 | 243,719 | 809 | 608 | 94,309 | 274,421 | 59,100 | 240,559 |
| | **M7CS** | **21** | 513 | -54,246 | 243,009 | **15** | 824 | 308 | 396 | 55,853 | 242,634 | **477** | 667 |
| | M7EH | 276,569 | 330,251 | 110,812 | 480,919 | 12,541 | 348,603 | 278,697 | 328,233 | 301,248 | 388,034 | 137,245 | 319,811 |
| | M7HT | 25,394 | 77,507 | 196,757 | 555,710 | 315,226 | 404,679 | 26,483 | 77,095 | 365,104 | 459,917 | 315,227 | 404,679 |
| | **M7RS** | 381,829 | 357,622 | **-13,893** | 642,573 | 240,287 | 503,510 | 381,829 | 357,622 | 430,257 | 471,898 | 418,487 | 364,103 |
| CBS | **ACC** | **1,719** | 8,433 | 1,544 | 16,316 | -2,865 | 6,521 | 5,809 | 6,040 | 10,005 | 12,873 | 5,023 | 4,998 |
| | CEDD | 4,695 | 16,375 | -2,421 | 22,378 | -1,427 | 9.032 | 13,024 | 10,106 | 11,303 | 19,375 | 6,529 | 6,305 |
| | CL | -4,066 | 12,149 | 15,887 | 89,294 | 920 | 7,152 | 9,307 | 8,291 | 22,584 | 87,797 | 5,042 | 5,083 |
| | **FCTH** | -16,115 | 53,730 | 8,195 | 48,200 | 733 | 2,815 | 44,261 | 31,228 | 15,061 | 46,465 | **1,717** | 2,332 |
| | JCD | -5,215 | 13,714 | 32,538 | 202,048 | -3,601 | 14,188 | 11,441 | 8,450 | 49,546 | 198,463 | 11,389 | 9,002 |
| | JCH | -4,528 | 12,150 | 9,977 | 64,864 | -2,531 | 7,801 | 10,074 | 7,518 | 18,305 | 62,969 | 6,057 | 5,446 |
| | M7CL | -6,524 | 14,503 | 695 | 33,675 | -2,526 | 9,477 | 13,067 | 8,131 | 24,754 | 22,455 | 6,683 | 7,097 |
| | **M7CS** | -4,815 | 8,987 | **181** | 9,527 | **590** | 14,747 | **5,612** | 8,451 | 5,265 | 7,892 | 8,980 | 11,609 |
| | **M7EH** | -9,124 | 32,083 | 546 | 3,733 | 5,349 | 38,699 | 16,972 | 28,265 | **2,203** | 3,042 | 25,728 | 29,070 |
| | M7HT | -6,185 | 3,918 | 4,090 | 45,322 | 5,340 | 52,299 | 6,185 | 3,918 | 26,460 | 36,759 | 34,011 | 29,658 |
| | M7RS | -7,068 | 10,792 | 24,432 | 111,824 | 3,075 | 30,420 | 9,694 | 8,182 | 75,753 | 84,948 | 16,583 | 25,530 |


[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[3] G. Blakowski and R. Steinmetz. A media synchronization survey: reference model, specification, and case studies. *IEEE J. on Selected Areas in Communications*, 14(1):5–35, 1996.

[4] M. Bober. MPEG-7 visual shape descriptors. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6):716–719, 2001.

[5] M. Broilo, G. Boato, and F. G. De Natale. Content-based synchronization for multiple photos galleries. In *IEEE Int. Conf. on Image Processing*, pages 1945–1948, 2012.

[6] N. Conci, F. D. Natale, and V. Mezaris. Synchronization of multi-user event media (SEM) at MediaEval 2014: Task description, datasets, and evaluation. In *MediaEval 2014 Workshop*, 2014.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision*, 60(2):91–110, 2004.

[8] M. Lux. Lire: open source image retrieval in java. In *ACM Int. Conf. on Multimedia*, pages 843–846, 2013.

[9] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[10] T. Ojala, M. Pietikäinen, and T. Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In *Computer Vision - ECCV 2000*, pages 404–420, 2000.

[11] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. of Computer Vision*, 42(3):145–175, 2001.

[12] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali. Cluster-based landmark and event detection for tagged photo collections. *IEEE MultiMedia*, 18(1):52–63, 2011.

[13] C. Papagiannopoulou and V. Mezaris. Concept-based image clustering and summarization of event-related image collections. In *ACM Int. W. on Human Centered Event Understanding from Multimedia*, pages 23–28, 2014.

[14] D. Pelleg, A. W. Moore, et al. X-means: Extending K-means with efficient estimation of the number of clusters. In *Int. Conf. on Machine Learning*, pages 727–734, 2000.

[15] G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Social event detection using multimodal clustering and integrating supervisory signals. In *ACM Int. Conf. on Multimedia Retrieval*, pages 23:1–23:8, 2012.

[16] T. Reuter and P. Cimiano. Event-based classification of social media streams. In *ACM Int. Conf. on Multimedia Retrieval*, pages 22:1–22:8, 2012.

[17] 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 *ACM Int. Conf. on Multimedia*, pages 397–406, 2014.

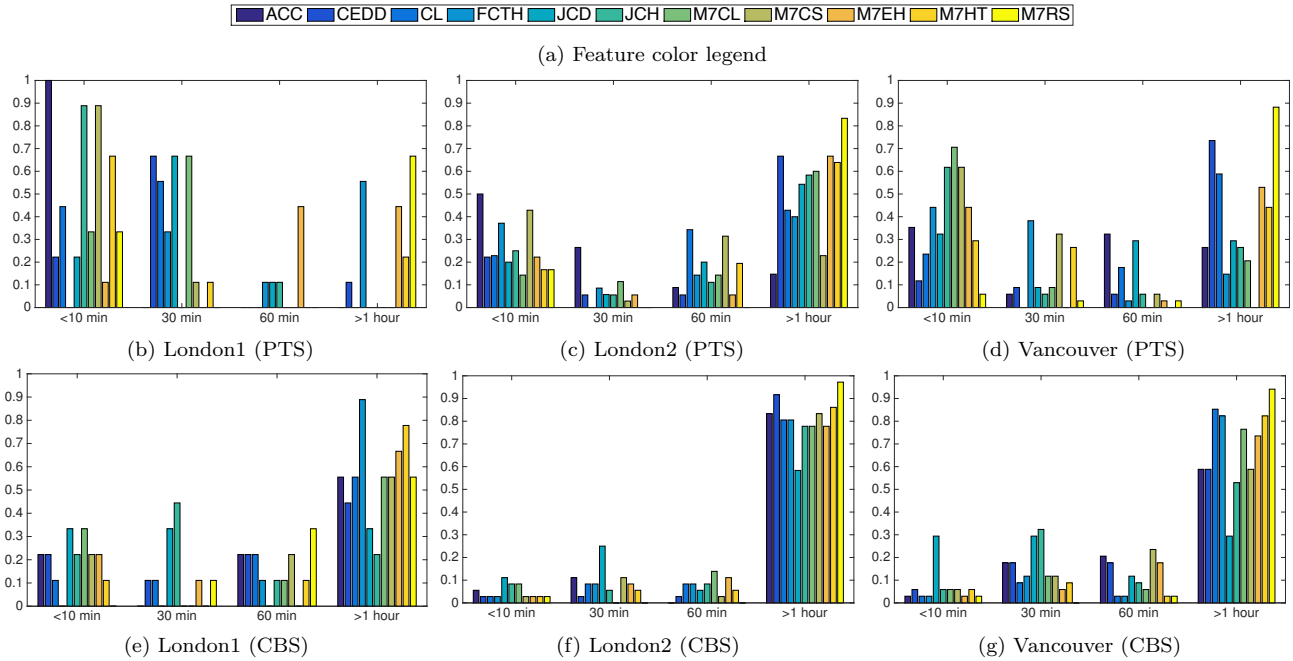
[18] E. Sansone, G. Boato, and M.-S. Dao. Synchronizing multi-user photo galleries with MRF. In *Media Synchronization Workshop*, 2014.

[19] A. Veenhuizen and R. van Brandenburg. Frame accurate media synchronization of heterogeneous media sources in an HBB context. In *Media Synchronization Workshop*, 2012.

[20] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3360–3367, 2010.

[21] J. Yang, J. Luo, J. Yu, and T. S. Huang. Photo stream alignment and summarization for collaborative photo collection and sharing. *IEEE Trans. on Multimedia*, 14(6):1642–1651, 2012.

[22] M. Zaharieva, M. Zeppelzauer, M. Del Fabro, and D. Schopfhauser. Social event mining in large photo collections. In *ACM Int. Conf. on Multimedia Retrieval*, pages 11–18, 2015.




(a) Feature color legend

(b) London1 (PTS) (c) London2 (PTS) (d) Vancouver (PTS)

(e) London1 (CBS) (f) London2 (CBS) (g) Vancouver (CBS)

Figure 3: Distribution of the time deviation (in minutes) between estimated and real time offsets for the three datasets. First row: PTS-based synchronization. Second row: CBS-based synchronization.

Table 4: Clustering results on sub-event detection. $C$ is the number of detected clusters. Bold values indicate best performance in terms of $F1$-score for each dataset and each clustering algorithm.

| Clustering | Feature | London1 | | | | | London2 | | | | | Vancouver | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | C | R | P | F1 | NMI | C | R | P | F1 | NMI | C | R | P | F1 | NMI |
| Visual (AHC) | ACC | 148 | 0.41 | 0.85 | 0.55 | 0.82 | **1083** | **0.30** | **0.88** | **0.45** | **0.81** | **682** | **0.21** | **0.88** | **0.34** | **0.76** |
| | CEDD | 129 | 0.43 | 0.70 | 0.54 | 0.78 | 1103 | 0.26 | 0.83 | 0.39 | 0.80 | 721 | 0.17 | 0.78 | 0.27 | 0.71 |
| | CL | 97 | 0.56 | 0.55 | 0.56 | 0.77 | 1373 | 0.18 | 0.83 | 0.39 | 0.80 | 888 | 0.11 | 0.85 | 0.19 | 0.72 |
| | FCTH | 138 | 0.40 | 0.76 | 0.52 | 0.80 | 1146 | 0.24 | 0.85 | 0.37 | 0.80 | 701 | 0.17 | 0.82 | 0.29 | 0.73 |
| | JCD | **113** | **0.50** | **0.66** | **0.57** | **0.79** | 1131 | 0.25 | 0.85 | 0.39 | 0.80 | 761 | 0.15 | 0.88 | 0.25 | 0.74 |
| | JCH | 130 | 0.42 | 0.71 | 0.53 | 0.79 | 1185 | 0.21 | 0.84 | 0.33 | 0.79 | 761 | 0.15 | 0.88 | 0.25 | 0.74 |
| | M7CL | 132 | 0.41 | 0.74 | 0.52 | 0.80 | 1167 | 0.20 | 0.80 | 0.32 | 0.78 | 844 | 0.11 | 0.82 | 0.19 | 0.71 |
| | M7CS | 134 | 0.42 | 0.81 | 0.56 | 0.83 | 966 | 0.28 | 0.84 | 0.42 | 0.81 | 648 | 0.17 | 0.88 | 0.28 | 0.76 |
| | M7EH | 139 | 0.39 | 0.64 | 0.48 | 0.71 | 1361 | 0.17 | 0.80 | 0.27 | 0.76 | 868 | 0.12 | 0.83 | 0.20 | 0.71 |
| | M7HT | 118 | 0.31 | 0.58 | 0.40 | 0.72 | 1117 | 0.16 | 0.72 | 0.26 | 0.75 | 659 | 0.13 | 0.72 | 0.21 | 0.70 |
| | M7RS | 194 | 0.26 | 0.79 | 0.40 | 0.77 | 1117 | 0.16 | 0.72 | 0.26 | 0.75 | 954 | 0.08 | 0.81 | 0.15 | 0.70 |
| Visual (X-Means) | ACC | 21 | 0.57 | 0.25 | 0.34 | 0.59 | 90 | 0.16 | 0.07 | 0.10 | 0.39 | 89 | 0.19 | 0.21 | 0.20 | 0.46 |
| | CEDD | 78 | 0.25 | 0.33 | 0.28 | 0.62 | 101 | 0.63 | 0.10 | 0.17 | 0.29 | 89 | 0.83 | 0.07 | 0.13 | 0.14 |
| | CL | 73 | 0.30 | 0.41 | 0.34 | 0.66 | **101** | **0.50** | **0.38** | **0.43** | **0.73** | **97** | **0.31** | **0.42** | **0.36** | **0.67** |
| | FCTH | 69 | 0.26 | 0.35 | 0.30 | 0.63 | 73 | 0.23 | 0.17 | 0.20 | 0.54 | 70 | 0.13 | 0.11 | 0.12 | 0.35 |
| | JCD | **69** | **0.46** | **0.58** | **0.51** | **0.78** | 89 | 0.46 | 0.29 | 0.36 | 0.68 | 89 | 0.31 | 0.34 | 0.32 | 0.61 |
| | JCH | 85 | 0.25 | 0.37 | 0.30 | 0.64 | 92 | 0.23 | 0.09 | 0.13 | 0.42 | 89 | 0.22 | 0.14 | 0.17 | 0.42 |
| | M7CL | 85 | 0.24 | 0.35 | 0.29 | 0.61 | 89 | 0.34 | 0.22 | 0.27 | 0.61 | 89 | 0.12 | 0.11 | 0.11 | 0.37 |
| | M7CS | 74 | 0.65 | 0.30 | 0.41 | 0.66 | 89 | 0.50 | 0.36 | 0.42 | 0.73 | 86 | 0.30 | 0.37 | 0.33 | 0.64 |
| | M7EH | 73 | 0.24 | 0.31 | 0.27 | 0.60 | 90 | 0.28 | 0.14 | 0.19 | 0.49 | 87 | 0.16 | 0.19 | 0.17 | 0.46 |
| | M7HT | 42 | 0.68 | 0.17 | 0.27 | 0.27 | 89 | 0.12 | 0.05 | 0.07 | 0.36 | 87 | 0.08 | 0.08 | 0.08 | 0.29 |
| | M7RS | 73 | 0.21 | 0.27 | 0.23 | 0.56 | 89 | 0.15 | 0.06 | 0.09 | 0.40 | 89 | 0.10 | 0.10 | 0.10 | 0.37 |
| Visual (AHC) & Time [early fusion] | ACC | **136** | **0.46** | **0.74** | **0.57** | **0.79** | 1081 | 0.26 | 0.86 | 0.40 | 0.81 | 748 | 0.17 | 0.85 | 0.28 | 0.74 |
| | CL | 111 | 0.32 | 0.53 | 0.40 | 0.71 | 1021 | 0.19 | 0.68 | 0.30 | 0.75 | 709 | 0.09 | 0.63 | 0.16 | 0.65 |
| | JCD | 117 | 0.35 | 0.64 | 0.45 | 0.75 | 957 | 0.22 | 0.72 | 0.33 | 0.77 | 712 | 0.10 | 0.69 | 0.18 | 0.68 |
| | M7CS | 118 | 0.41 | 0.73 | 0.53 | 0.80 | **660** | **0.29** | **0.80** | **0.53** | **0.84** | **542** | **0.18** | **0.77** | **0.29** | **0.74** |
| Visual (X-Means) & Time [early fusion] | ACC | 84 | 0.24 | 0.34 | 0.28 | 0.61 | 95 | 0.28 | 0.16 | 0.21 | 0.52 | 94 | 0.16 | 0.20 | 0.18 | 0.48 |
| | CL | 85 | 0.26 | 0.38 | 0.31 | 0.65 | 100 | 0.48 | 0.38 | 0.42 | 0.73 | 98 | 0.30 | 0.40 | 0.34 | 0.66 |
| | JCD | **73** | **0.50** | **0.69** | **0.58** | **0.82** | **97** | **0.60** | **0.42** | **0.49** | **0.79** | 80 | 0.26 | 0.30 | 0.28 | 0.62 |
| | M7CS | 84 | 0.26 | 0.37 | 0.30 | 0.65 | 99 | 0.49 | 0.38 | 0.43 | 0.74 | **97** | **0.30** | **0.41** | **0.35** | **0.67** |
| Visual (AHC) & Time [22] | ACC | 48 | 0.52 | 0.24 | 0.33 | 0.45 | 555 | 0.40 | 0.37 | 0.39 | 0.60 | 380 | 0.35 | 0.34 | 0.35 | 0.45 |
| | CL | 29 | 0.79 | 0.15 | 0.26 | 0.23 | 442 | 0.60 | 0.30 | 0.40 | 0.48 | 317 | 0.46 | 0.33 | 0.38 | 0.43 |
| | JCD | 41 | 0.73 | 0.19 | 0.30 | 0.31 | 617 | 0.31 | 0.41 | 0.36 | 0.62 | 271 | 0.52 | 0.27 | 0.36 | 0.39 |
| | M7CS | **73** | **0.72** | **0.66** | **0.68** | **0.82** | **383** | **0.62** | **0.67** | **0.65** | **0.85** | 298 | **0.53** | **0.63** | **0.58** | **0.78** |
| Visual (AHC) & Location [22] | ACC | 116 | 0.50 | 0.74 | 0.59 | 0.82 | 773 | 0.46 | 0.80 | 0.59 | 0.85 | 422 | 0.45 | 0.68 | 0.54 | 0.75 |
| | CL | 78 | 0.61 | 0.47 | 0.53 | 0.72 | 974 | 0.32 | 0.73 | 0.44 | 0.79 | 552 | 0.29 | 0.61 | 0.40 | 0.66 |
| | JCD | 104 | 0.57 | 0.62 | 0.57 | 0.78 | 869 | 0.41 | 0.73 | 0.52 | 0.81 | 427 | 0.44 | 0.50 | 0.47 | 0.58 |
| | M7CS | **106** | **0.53** | **0.72** | **0.61** | **0.83** | **635** | **0.53** | **0.71** | **0.61** | **0.84** | **349** | **0.43** | **0.57** | **0.55** | **0.72** |

# Paper V

# Multimodal Synchronization of Image Galleries

# Multimodal Synchronization of Image Galleries

Maia Zaharieva[1,2] Michael Riegler[3] Manfred Del Fabro[4]

[1]Interactive Media Systems Group, Vienna University of Technology, Austria
[2]Multimedia Information Systems Group, University of Vienna, Austria
[3]Media Performance Group, Simula Research Laboratory AS, Norway
[4]Distributed Multimedia Systems Group, Klagenfurt University, Austria
maia.zaharieva@tuwien.ac.at, michael@simula.no, manfred.delfabro@aau.at

## ABSTRACT

This paper describes our contribution to the MediaEval 2014 task on the Synchronization of multi-user Event Media (SEM). We propose two multimodal approaches that employ both visual and time information for the synchronization of different images galleries and for the detections of sub-events. The methods prove robustness in the determination of time offsets with accuracy of up to 87%.

## 1. INTRODUCTION

A multifaceted view of a social event can emerge when different people capture different perspectives of the same event and a compilation of all images is created. While it is typically easy to get an overview of a single image gallery, it is much more difficult to synchronize the content of two or more collections. In general, there is no guarantee that timestamps, location information or textual descriptions associated with images are correct.

In our contributions to the SEM task [1] we first focus on global visual features to identify highly similar images across different galleries of the dataset. Following, we apply visual- and time-based methods for the synchronization of galleries and for the detection of sub-events. Our first approach relies on the pairwise comparison of images in order to link different galleries. Agglomerative Hierarchical Clustering (AHC) is applied in order to group image pairs to sub-events. The synchronization offsets are calculated by iterating through the image pairs in a transitive way. In our second approach all images are clustered using the XMeans algorithm in order to identify sub-events. The synchronization offsets are estimated by calculating average time differences within the clusters.

## 2. APPROACHES

### 2.1 AHC-based Approach

We employ AHC for both time offset calculation and sub-event detection. We first cluster all images of the dataset using the MPEG7 Color Structure Descriptor (MPEG7-CS). At the very lowest hierarchy level clusters of visually highly similar images are generated. We sort these pairs of images in ascending order according to their dissimilarity level. We consider such pairs of images identical if: 1) the images originate from different galleries and 2) the dissimilarity distance does not exceed a predefined threshold. Images, representing different galleries, are considered as entry points for the synchronization of the corresponding gallery. We process the sorted pair list until we are able to build a transitive list of entry points for all galleries presented in the full dataset or we reach the end of the list. Eventually, all galleries are time aligned according to the provided reference collection using the corresponding entry points.

A higher hierarchy level of AHC already provides a reliable base for visual-based detection of sub-events. In order to avoid the building of broad clusters, we employ a strict cutoff threshold in combination with the Ward method [4] to automatically define the number of clusters. We reduce the resulting over-segmentation of underlying events by employing an adaptive, time-based approach for cluster merging. Two clusters are merged if they share a common gallery and the minimum time distance between the corresponding images is lower than a predefined threshold.

### 2.2 XMeans-based Approach

For this approach we employ a modified version of the algorithm presented in [3]. We select the best global feature for the given dataset by considering the information gain. The calculation is done for 13 different features (Color and Edge Directivity Desciptor (CEDD), Fuzzy Color and Texture Histogram (FCTH), Joint Composite Descriptor (JCD), Pyramid Histogram of Ortented Gradients (PHOG), Edge Histogram (EH), Color Layout (CL), Gabor, Tamura, Luminance Layout (LL), Opponent Histogram (OH), JPEG Coefficent Histrogram (JPEGCoeff), Scaleable Color (SC) and Auto Color Correlogram (ACC) [2]). JCD had the highest information gain for the SEM dataset and, therefore, it was employed for this approach.

In order to synchronize the dataset, we first cluster all images using the XMeans algorithm. Following, we consider the average deviation of the reference image timestamps to all other images of a collection that share a common cluster as offset for this image collection. If there are less than two reference images in a cluster, we use the available corrected timestamp of non-reference images which already have an offset from another cluster. For sub-event detection, we employ XMeans clustering using JCD or the corrected capture times as features.

## 3. EXPERIMENTS AND RESULTS

The SEM development dataset contains 304 Flickr images from the *London Olympic Games 2012*. The images are arranged in 10 galleries and represent 59 sub-events in total.


*MediaEval 2014 Workshop,* October 16-17, 2014, Barcelona, Spain

**Table 1: Sub-event detection results on the development dataset in terms of number of detected clusters (C), F1-score (F1), and Normalized Mutual Information (NMI).**

| | C | F1 | NMI |
|---|---|---|---|
| Time-based clustering | 98 | 0.6363 | 0.8696 |
| AHC + MPEG7-CS | 91 | 0.5543 | 0.8179 |
| AHC + MPEG7-CS + Time | 45 | 0.6303 | 0.7927 |
| Xmeans + JCD | 89 | 0.5123 | 0.7812 |
| Xmeans + Time | 100 | 0.5731 | 0.8231 |

**Table 2: Official runs configurations.**

| | Time Offset | Sub-events detection |
|---|---|---|
| run 1 | AHC + MPEG7-SC | AHC + MPEG-7 SC |
| run 2 | AHC + MPEG7-SC | Time-based |
| run 3 | XMeans + JCD | XMeans + JCD |
| run 4 | XMeans + JCD | XMeans + Time |
| run 5 | AHC + MPEG7-SC | XMeans + Time |

Experiments on the development dataset show significant differences in the precision of detected time offsets between the two approaches. While, the AHC-based approach in combination with MPEG7-CS achieves 18.5 seconds deviation in average over the 10 galleries, the XMeans-based and the JCD feature obtain only 2216.4 seconds in average.

Additionally, we compare the performances of purely time-based clustering (after considering the time offsets), visual-based clustering, and the combination thereof using the AHC approach. We measure the performance by means of harmonic mean (F1-score) of recall and precision and Normalized Mutual Information (NMI) measuring the goodness of clustering of retrieved events. The results achieved show that both the time-based and the visual-based clustering result in over-segmentation of the underlying events (90+ detected sub-events vs. 59 ground truth events) and high NMI scores. The combination of visual and time information outperforms the visual-based approach and significantly reduces the number of detected sub-event clusters (see Table 1). Noteworthy is the observation that with both approaches, the time-based detection of sub-events outperforms the corresponding visual-based approach in terms of F1 (at higher over-segmentation costs).

We submitted five runs for the final evaluation (see Table 2 for the configurations). Tables 3 and 4 summarize the corresponding results for the synchronization and for the sub-event detection task. Results on the synchronization task are reported in terms of precision (percentage of synchronized galleries with a misalignment lower than 30 minutes), and accuracy (closeness of detected offset to real offset, normalized with respect to the maximum accepted time lapse of 30 minutes). The results achieved confirm our experiments on the development dataset: the AHC-based approach in combination with the MPEG-7-CS clearly outperform our XMeans-based approach. Although both datasets contain approximately the same number of galleries (35 Vancouver, 37 London) they perform differently. The Vancouver dataset was highly successfully aligned within the maximum accepted time lapse of 30 minutes with a precision of 94%. By contrast, the London dataset achieves a good overall performance by means of an accuracy of 87% at a significantly lower precision level of 47%. The results on the sub-event detection task are ambiguous. Overall, the AHC-based approach tends to detect a significantly larger number of sub-events than the XMeans-based approach. Nevertheless, both approaches result in high Random Index (RI) scores which reflects the purity of the detected clusters. While in general high RI scores may also be the result of strong over-segmentation, the number of detected clusters with our runs differ significantly.

**Table 3: *MediaEval 2014 Benchmark* results for the synchronization task in terms of precision (P) and accuracy (A).**

| | Vancouver dataset | | London dataset | |
|---|---|---|---|---|
| | P | A | P | A |
| AHC + MPEG7-SC | 0.9412 | 0.7919 | 0.4722 | 0.8746 |
| XMeans + JCD | 0.5882 | 0.5701 | 0.3611 | 0.4676 |

**Table 4: *MediaEval 2014 Benchmark* results for the sub-event detection task in terms of number of detected clusters (C), Random Index (RI), and F1-score (F1).**

| | Vancouver dataset | | | London dataset | | |
|---|---|---|---|---|---|---|
| | C | RI | F1 | C | RI | F1 |
| run 1 | 379 | 0.9787 | 0.1012 | 368 | 0.9842 | 0.2614 |
| run 2 | 709 | 0.9782 | 0.0505 | 709 | 0.9873 | 0.1687 |
| run 3 | 91 | 0.9610 | 0.1087 | 91 | 0.9760 | 0.1331 |
| run 4 | 81 | 0.9687 | 0.0890 | 81 | 0.9797 | 0.1653 |
| run 5 | 98 | 0.9727 | 0.1079 | 98 | 0.9797 | 0.1653 |

# 4. CONCLUSION

In this paper we presented two multimodal approaches for the synchronization of multi-user galleries and for the detection of sub-events. The results obtained on the SEM datasets indicate the potential of the combination of visual and time information for the tasks. An open issue is the detection of sub-events that are visually highly similar and that take place in a short time period.

## Acknowledgments

This work has been partly funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-010, by the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214/22573/33955, and by the iAD center for Research-based Innovation (project number 174867) funded by the Norwegian Research Council.

# 5. REFERENCES

[1] N. Conci, F. D. Natale, and V. Mezaris. Synchronization of multi-user event media (SEM) at MediaEval 2014: Task description, datasets, and evaluation. In *MediaEval 2014 Workshop*, 2014.

[2] M. Lux. Lire: Open source image retrieval in java. In *ACM Int. Conf. on Multimedia*, pages 843–846, 2013.

[3] M. Riegler, M. Lux, and C. Kofler. Frame the crowd: Global visual features labeling boosted with crowdsourcing information. In *MediaEval 2013 Workshop*, 2013.

[4] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

# Paper VI

# Introduction to a Task on Context of Experience: Recommending Videos Suiting a Watching Situation

# Introduction to a Task on Context of Experience: Recommending Videos Suiting a Watching Situation

Michael Riegler[1], Martha Larson[3], Concetto Spampinato[2], Jonas Markussen[1]
Pål Halvorsen[1], Carsten Griwodz[1]
[1]Simula Research Laboratory, Norway
[2]University of Catania, Italy
[3]Delft University of Technology, Netherlands
{michael, jonassm, paalh, griff}@simula.no, cspampin@dieei.unict.it, m.a.larson@tudelft.nl

## ABSTRACT

We propose a Context of Experience task, whose aim it is to explore the suitability of video content for watching in certain situations. Specifically, we look at the situation of watching movies on an airplane. As a viewing context, airplanes are characterized by small screens and distracting viewing conditions. We assume that movies have properties that make them more or less suitable to this context. We are interested in developing systems that are able to reproduce a general judgment of viewers about whether a given movie is a good movie to watch during a flight. We provide a data set including a list of movies and human judgments concerning their suitability for airplanes. The goal of the task is to use movie metadata and audio-visual features extracted from movie trailers in order to automatically reproduce these judgments. A basic classification system demonstrates the feasibility and viability of the task.

## 1. INTRODUCTION

The challenge of the Context of Experience task is to automatically predict viewers' judgments on whether video content is suitable for a particular watching situation. Ultimately, the aim is to build a recommender system that would provide viewers with recommendations of content for a given context. Currently, the majority of work on video content recommendation focuses on personal preferences, and overlooks cases in which context might have a strong impact on preference relatively independently of the personal tastes of specific viewers. Particularly strong influence of context can be expected in psychologically stressful or physically uncomfortable situations.

For our task, we choose one such situation, with which a large number of people have quite frequent experience: watching movies on an airplane. In this situation, a large majority of viewers share a common goal, which we consider to be a *viewing intent*. The goal is to pass time as pleasantly and meaningfully as possible, while confined in the small space of an airplane cabin, which is characterized by a number of distractors. We take the large number of websites discussing movies to watch on airplanes (e.g., [9]) as evidence that this viewing intent is dominant among air travellers. Although the scope of this task is limited to the airplane scenario, we emphasize that the challenge of Context of Experience is a much broader area of interest. Other examples of stressful contexts where videos are becoming increasingly important include hospital waiting rooms, and dentists offices, where videos are shown during treatment.

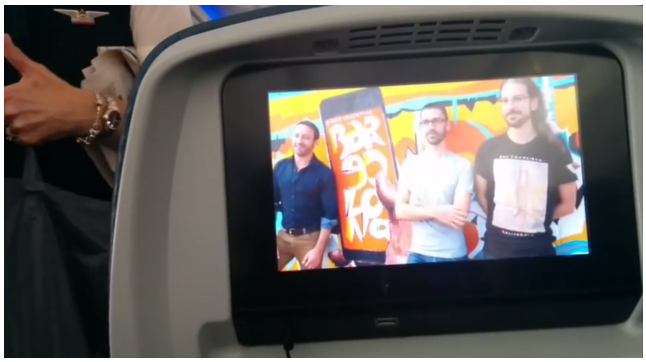


**Figure 1: A set of conditions, including small screen and confined, crowded space, characterize the context of watching a movie on an airplane.**

## 2. TASK DESCRIPTION

For the task we provide the participants a list of movies, including links to descriptions and video trailers. The assignment of the task is to classify each movie into +goodonairplane / -goodonairplane classes. Therefore, the ground truth of the task is derived from two sources: A list of movies actually used by a major airline[1], as well as user judgments on movies that are collected via a crowdsourcing tool[2]. Task participants should form their own hypothesis about what is important for users viewing movies on an airplane, and design an approach using appropriate features and a classifier, or decision function. Figure 1 gives an impression of a screen commonly used on an airplane and the very specific attributes regarding size and quality of the video. The value of the task lies in understanding the ability of content-based and metadata-based features to discriminate the kind of movies that people would like to watch on small screens under stressful or somehow not normal situations. Since the multimedia content that users watch on flights can influence their well being and overall experience this task is related to the quality of multimedia experience work like for exam-

[1]http://www.klm.com/travel/no_en/prepare_for_travel/on_board/entertainment/onboard_movies.htm
[2]https://crowdflower.com/


*MediaEval 2015 Workshop,* Sept. 14-15, 2015, Wurzen, Germany

ple [6, 5, 3, 4, 1]. Apart from that, the task also includes the area of user intent since the intent of the users, why they want to watch movies on the airplane, is a strong influencing force on what they watch [7, 8]. Task participants are provided with a collection of videos, i.e., trailers as a representative for the movie because of copyright issues, and the context, e.g., video URL, metadata, user votes etc. Apart from that we also provide different pre-extracted features, including visual and audio features. The participants are asked to develop methods that will predict to which intent class the video belongs, respectively, good or bad to watch on an airplane.

To tackle the task it can be addressed by leveraging techniques from multiple multimedia-related disciplines, including such as social computing (intent), machine learning (classification), multimedia content analysis, multimodal fusion, and crowdsourcing. Further we hope that it will be useful for content provider, since the exploitation of intent in combination with users' satisfaction could lead to more sophisticated ways to develop methods of providing a better service to the users.

## 3. DATA SET

The data set we provide is released, including titles and links, that allow participants to gather online metadata and trailers for movies. We do not, as already mentioned, provide the video files because of copyright restrictions. Movies are collected based on movie lists from a major international airline, in our case, KLM Royal Dutch Airlines. The final list of movies is a merged set of movies collected between February and April 2015. The video data set contains both positive and negative samples, whereas the negative examples are carefully sampled from IMDB in order to create a fair and representative negative class. The data set is split into a training set and a test set. In order to collect user judgments, we use an existing system that has been built for the purpose of collecting user feedback of this sort. We evaluate systems both with respect to the airline's choice of movies, and the crowd's choice of airline-suitable movies. Votes about the labels collected by crowdsourcing are considered as the authoritative labels. For this reason, crowdworkers are asked to rank a small set of movies with respect to how strongly they would like to watch this video on an airplane. This ranking is then combined to create the class for each movie in the training and test data.

**Technical details.** Overall, the data set contains 318 movies. Links to trailers are collected from IMDB and YouTube. Participants are also allowed to collect their own data such as full length movies, more metadata and user comments, etc. The goal of systems that are developed to address this task should be to automatically identify appropriate content, i.e., whether a movie should be recommended for being watched on an airplane or not. To achieve this goal, the methods should not require manual or crowdsourced input. The data set contains extracted visual, audio and text features. Furthermore, we provide metadata collected from IMDB including user comments. The visual features that are provided are: Histogram of Oriented Gradients (HOG), Color Moments, Local Binary Patterns (LBP) and Gray Level Run Length Matrix. The audio descriptors are Mel-Frequency Cepstral Coefficients (MFCC). The development set contains 95 labelled movies. The test data contains 223 movies without labels.

| Features used | Precision | Recall | F1-score |
|---|---|---|---|
| Metadata + user ratings | 0.581 | 0.6 | 0.583 |
| Only user ratings | 0.371 | 0.609 | 0.461 |
| Only visual information | 0.447 | 0.476 | 0.458 |
| Only metadata | 0.524 | 0.516 | 0.519 |

**Table 1: Classification in terms of weighted average of precision, recall and F1-score.**

**Evaluation.** For the evaluation we use the standard metrics Precision, Recall and weighted F1 score. Negative and positive classes in both data sets are balanced. Participants are asked to submit a predicted class for each movie in the test data set. The metrics then are calculated and provided to the participants. For a transparent and fair procedure, the labels used for the evaluation will be released together with the results.

**Initial results.** To confirm the viability of the task, and show the possibilities opened by this data set, we carried out some basic classification experiments. For the classification we used the Weka library. As classifier we choose the rule based PART classifier. This classifier uses separate and conquer to generate a decision list. From this, it builds a decision tree where the best leaves are used as rules for the classifier [2]. Table 1 show the results of our four initial runs. For the evaluation we used the weighted average of precision, recall and F1-score. The first run uses metadata (language, year published, genre, country, runtime and age rating) in combination with user ratings as input for the classifier. This run is our best performer. It clearly outperforms the naive baseline, which is 0.5 (precision, recall and F1-score). The second run uses user ratings alone. This run performs well with recall, but poorly with precision. This implies that receiving certain user ratings is a necessary, but not a sufficient condition for being a movie that is good to watch on an airplane. Taken together, the first two runs confirm that the task is non-trivial, and that it is also viable. The third run uses visual features. This run scores below the naive baseline. However, the approach to visual classification here was relatively simple. Additional exploratory experiments, not reported here, revealed that visual features do have the ability to approve results when used in combination with other features. Such combinations are interesting for future work.

Finally, the last run confirms that metadata without user ratings is able to yield performance above the naive baseline. An information gain based analysis of all features ranked genre, publication year, country, language and runtime as the top five features.

## 4. SUMMARY

The task is challenging due to the complex relationship between the multimedia content, and viewers' perceptions and reception. We hope that the novel use case will inspire researchers to investigation of user intent and context of experience. Understanding user intent and what users need in order to have the best experience is an important emerging topic in the area of multimedia research.

## 5. ACKNOWLEDGMENT

This work is funded by the Norwegian FRINATEK project "EONS" (#231687) & the EC project CrowdRec (#610594).

## 6. REFERENCES


[1] A. Borowiak and U. Reiter. Long duration audiovisual content: Impact of content type and impairment appearance on user quality expectations over time. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 200–205. IEEE, 2013.

[2] E. Frank and I. H. Witten. Generating accurate rule sets without global optimization. 1998.

[3] P. Lebreton, A. Raake, M. Barkowsky, and P. Le Callet. Evaluating complex scales through subjective ranking. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pages 303–308. IEEE, 2014.

[4] B. Rainer and C. Timmerer. A quality of experience model for adaptive media playout. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pages 177–182. IEEE, 2014.

[5] J. A. Redi, Y. Zhu, H. de Ridder, and I. Heynderickx. How passive image viewers became active multimedia users. In *Visual Signal Quality Assessment*, pages 31–72. Springer, 2015.

[6] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank. Factors influencing quality of experience. In *Quality of Experience*, pages 55–72. Springer, 2014.

[7] M. Riegler, L. Calvet, A. Calvet, P. Halvorsen, and C. Griwodz. Exploitation of producer intent in relation to bandwidth and qoe for online video streaming services. In *Proceedings of the 25th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 7–12. ACM, 2015.

[8] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. 2010.

[9] Tripinsurance. Best movies guide for airplanes. http://www.tripinsurance.com/tips/guide-to-the-best-moviestv-shows-to-watch-on-a-plane. [last visited, Dezember. 10, 2014].

# Paper VII

# Right inflight? A Dataset for Exploring the Automatic Prediction of Movies Suitable for a Watching Situation

# Right Inflight? A Dataset for Exploring the Automatic Prediction of Movies Suitable for a Watching Situation

Michael Riegler[1], Martha Larson[2], Concetto Spampinato[3], Pål Halvorsen[1], Mathias Lux[4]
Jonas Markussen[1], Konstantin Pogorelov[1], Carsten Griwodz[1], Håkon Stensland[1]
[1]Simula Research Laboratory & University of Oslo
[2]TU Delft & Radboud University Nijmegen
[3]University of Catania
[4]University of Klagenfurt
michael@simula.no, m.a.larson@tudelft.nl, cspampin@dieei.unict.it, mlux@itec.aau.at

## ABSTRACT

In this paper, we present the dataset *Right Inflight* developed to support the exploration of the match between video content and the situation in which that content is watched. Specifically, we look at videos that are suitable to be watched on an airplane, where the main assumption is that that viewers watch movies with the intent of relaxing themselves and letting time pass quickly, despite the inconvenience and discomfort of flight. The aim of the dataset is to support the development of recommender systems, as well as computer vision and multimedia retrieval algorithms capable of automatically predicting which videos are suitable for inflight consumption. Our ultimate goal is to promote a deeper understanding of how people experience video content, and of how technology can support people in finding or selecting video content that supports them in regulating their internal states in certain situations. *Right Inflight* consists of 318 human-annotated movies, for which we provide links to trailers, a set of pre-computed low-level visual, audio and text features as well as user ratings. The annotation was performed by crowdsourcing workers, who were asked to judge the appropriateness of movies for inflight consumption.

## CCS Concepts

**•Information systems → Information retrieval; Multimedia and multimodal retrieval;**

## Keywords

Multimedia; Intent; Context; Data Set

## 1. INTRODUCTION

Increasingly, researchers are interested in developing multimedia analysis techniques that can predict the affective impact of video on viewers, and in releasing datasets that will support this work [26, 2]. Such work focuses on the way that viewers experience video. It fails, however, to take into account that a viewer does not view a video in a vacuum. Rather, viewing a video involves simultaneously experiencing *the video* and also *the context in which the video is viewed*. These two experiences interact, giving rise to a new challenge for multimedia research, which we call *Context of Experience*. Two major considerations underlie the importance of Context of Experience. First, we anticipate that techniques that are able to determine the suitability of videos for particular contexts of experience would be applicable for a wide range of users. For cases in which context has a strong impact on how a viewer experiences video, we expect that context could be an important predictor of viewer preference for videos, overshadowing personal taste or mood. Second, Context of Experience is closely related to user viewing intent, i.e., the reason why a person is watching a particular video. If we are able to predict the suitability of a video for a context, we are able to give viewers more useful tools for finding and selecting content that will help them in a given situation, for example, self-regulating in a situation of psychological tension.

In this paper, we focus on the case of viewers watching movies on an airplane. Here, independently of personal preferences, viewers share the common goal, which we consider to be a *viewing intent*, of passing time and keeping themselves occupied by being entertained while being confined in an uncomfortable small space of an airplane cabin. The dataset is designed to help answer the question of whether it

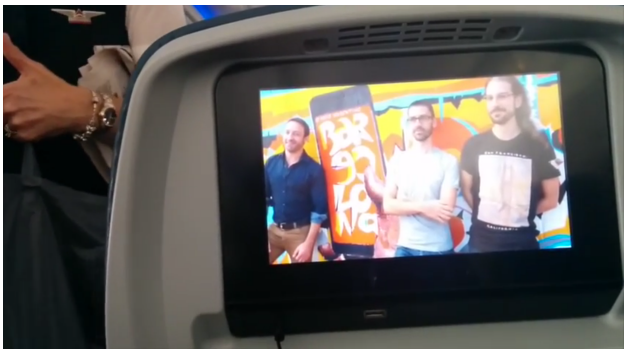


**Figure 1: A set of conditions, including small screen and confined, crowded space, characterize the context of watching a movie on an airplane.**



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



ISBN 978-1-4503-4297-1/16/05…$15.00

DOI: http://dx.doi.org/10.1145/2910017.2910619

is possible to predict which movies allow viewers to achieve the goal of passing time, relax or distracting themselves, given the context. Furthermore, the context of airline travel includes also limitations of the employed technology (e.g., screen size) and the environment itself (e.g., background noise, interruptions, presence of strangers). We have chosen the airplane scenario as the role of stress and viewers' intent to distract themselves is widely acknowledged [29]. Figure 1 gives an impression of a screen commonly used on an airplane and the very specific attributes regarding size and quality of the video.

Although the scope of the proposed dataset is limited to the airplane scenario, the challenge of Context of Experience is a much broader area of interest. Other examples of stressful contexts where videos are becoming increasingly important include hospital waiting rooms, and dentists offices where videos are shown during treatment. The dataset was initially developed, but not public released, as part of the MediaEval Multimedia Benchmark [12, 7] and a preliminary description can be found in [18].

The dataset comprises a list of 318 movies, including links (movies or trailers are not provided because of copyright issues) to descriptions and video trailers, as well as a set of pre-extracted visual, audio and text features from movies, along with annotations created by human judges, in this case, crowdsourcing workers. Additional information about the context such as metadata and user votes/rates is also given.

The dataset is designed to support binary classification of movies as either `+goodonairplane` or the `-goodonairplane` class. For this reason, the ground truth of the task is derived from two sources: A list of movies actually used by a major airline [6], as well as user judgments on movies that are collected via a crowdsourcing platform [4].

In order to address the Context of Experience challenge instantiated by this dataset, researchers can form their own hypothesis to find out what is important for users. This can be done for example by using most appropriate low-level features extracted from airplane's movies, and, accordingly, design approaches using appropriate features, classifiers, recommender system or decision function. The value of the dataset lies in understanding the ability of content-based and metadata-based features to discriminate the kind of movies that people would like to watch on small screens under stressful or somehow not normal situations. The *Right Inflight* dataset can be addressed with a variety of multimedia-related methods, like for example, recommender systems, social computing (intent), machine learning (classification), multimedia content analysis, multimodal fusion and crowdsourcing.

Further, we hope that the insights that can be gained with this dataset will be useful for content providers. If it is possible to understand how user intent contributes to user satisfaction, it would be possible to provide users with more sophisticated content recommendation and delivery services.

## 2. RELATED WORK

The challenge of Context of Experience stands at the intersection of research efforts currently ongoing in two different disciplines. First, in the field of multimedia, it is related to work on the impact of video content on viewers. Several datasets and benchmarks have contributed to supporting research that develops algorithms capable of automatically predicting the emotional impact (affective impact) of video content on the viewer. Within the MediaEval benchmark [12], these have been an early task on predicting viewer experienced boredom [26] and a current task on the affective impact of movies [25]. Moreover, in the field of multimedia, extensive work has been carried out on Quality of Experience, including [16, 15, 8, 14, 3]. Finally, Context of Experience is related to multimedia research in the area of viewer intent [17], since the intent of users (i.e., the reason why they want to watch movies on the airplane) is a strong influencing force on what they watch [17].

Second, in the field of recommender systems, Context of Experience is related to work on context-aware recommendation [1, 24]. Researchers have devoted significant effort into organizing challenges in the area of context-aware movie recommendation [22, 23]. There is, however, a critical difference between the challenge of Context of Experience and the challenge of context-aware movie recommendation. Context of Experience assumes that the experience of viewing a movie interacts with the context in which a movie is viewed. As a result, the movie is actually able to *change* the context. By conceptualizing context as Context of Experience we focus on the possibility that viewers might choose to view a movie driven by a particular intent, i.e., a goal. In the context of airline travel, which we assume has a strong interaction with the movie viewing experience, we assume the goal of the viewers is to be more comfortable and past time. Addressing Context of Experience means that we are not 'just' matching movies with personal tastes, but actually helping users accomplish goals. Although, personal preference without doubt plays a key role in determining which movies that people most enjoy during air travel, it is important that recommender systems are also able to exploit the general, context-related, tendency for people to find certain movies more suitable than others for watching on an airplane.

Datasets for research in computer science are an important tool to allow researchers to exchange and compare methods, techniques and algorithms. In information retrieval, large collections of document are used to evaluate for instance new ranking mechanisms or relevance functions. Due to the ever-changing nature of available data, new datasets are necessary. Recently there has been a move to develop datasets that consist of Creative Commons material. This movement helps the community to overcome the challenge of dealing with licensing restrictions, which effectively limit both the collection and the redistribution of data. Some datasets are released with the idea that they will be used for multiple purposes, for example, YFCC100M, a large-scale Flickr image dataset [28]. Whereas other datasets are released with annotations.

A key example is the LIRIS-ACCEDE (Annotated Creative Commons Emotional DatabasE) dataset for affective video content analysis [2], already mentioned in the Introduction.

Across the areas of multimedia and recommender systems, it is notable that few datasets focused on the actual intent of the users and the context. To the best of our knowledge, there is only one dataset including multimedia data (images in this case) as well as the photographers' intent, namely [11]. To create this dataset, photographers on Flicker were asked for permission to include their photo in the dataset as well as to take part in a survey, which aimed

at uncovering the actual reason why the people took the photo. Possible answers ranged from to publish it online, to capture a moment, to preserve a feeling. The data then was double checked in an evaluation run on Amazon Mechanical Turk. Both the photo survey as well as the results from Mechanical Turk are part of the dataset.

## 3. DATA COLLECTION

The dataset was collected in a series of steps. First, we collected the names of all the movies that were shown on flights by KLM between February 2015 and April 2015 from the KLM website [6]. We ended up with 201 movies for February, 196 for March and 200 for April. The movies were also ordered into 7 categories by KLM. The categories were *Latest, Recent, The collection, Family, World, Dutch movies* and *European movies*. Some of the movies appeared several times in different months. In the final list of movies, each movie only appeared once. The selection of movies that we included in the dataset contained 318 movies containing videos collected from KLM as positive examples and carefully selected negative examples from movie databases. For negative examples, we chose movies of the same categories and released around the same time of positive samples, but not used in the KLM system.

For the movies in this list, we crawled (i) metadata from popular movie ranking websites like IMDb and Rotten Tomatoes, etc. and (ii) links to movie trailers and posters. Afterwards, we conducted a crowdsourcing study using the Crowdflower [4] platform in two steps. First, we asked the study participants about their flying experience and their experience with movies in order to identify crowdworkers (people who do tasks on crowdsourcing platforms) who had watched movies during a flight. When we collected a large enough subset of flight experienced workers, we performed a second study.

In the second part, we asked the workers to rank the movies of our first collected list in terms of how likely they would watch the movies during a flight. This study is described in more detail in the next section.

## 4. CROWDSOURCING OF MOVIE PREFERENCES

Since crowdsourcing of subjective information is quite challenging, we followed the principles discussed in [30] and [19]. In our crowdsourcing study, we collected opinions concerning whether people would like to watch a movie on an airplane or not. Each worker was given 3 trailers to watch plus a short video intended to help them recall the situation of being on an airplane [1]. Figure 3 shows the task description presented to the crowdsourcing workers. After they looked at the trailers, we asked some questions. First, we asked them to provide us the title of each movie in order to check whether the crowdworkers actually watched the movies or just rushed trough the questions. After that, we asked them to rank the videos from 1 to 3 according to the likelihood (1 the most likely, 3 the least likely) they would watch those videos during a flight. Crowdworkers were also asked to provide a short explanation/motivation of their ranking as well as their favorite movie genre. For each movie, we collected at least five rankings from different users. From these

[1]https://youtu.be/TxC3OV9dBeo


**Figure 2: Task presentation to the workers. Each worker was given 3 movies to rank and some additional questions to answer.**

| Run | Features used | Precision | Recall | F1-score |
|---|---|---|---|---|
| i | Metadata + user ratings | 0.581 | 0.6 | 0.583 |
| ii | Only user ratings | 0.371 | 0.609 | 0.461 |
| iii | Metadata + Visual | 0.584 | 0.6 | 0.586 |
| iv | Only visual information | 0.447 | 0.476 | 0.458 |
| v | Only metadata | 0.524 | 0.516 | 0.519 |

**Table 1: Classification in terms of weighted average of precision, recall and F1-score for different types of input data used.**

rankings, we calculated the average rank that was used to determine the label `+goodonairplane` or `-goodonairplane`. For movies, for which we could not make a clear decision, we collected more crowdsourcing data to break the tie.

All in all, we had 548 different workers participating in the task who provided 1644 judgments. From these 1644 judgments, we used 1590 after discarding workers who provided answers that clearly reflected that they did not take the task seriously. To detect such non-serious workers, we checked over the provided movie titles and questionnaire completion times. A very fast finishing time was defined as faster as the average of three people in our laboratory could read and finish the task if they tried to do it very fast. Which was circa 3 minutes plus the time of the trailers. We discarded around 20% of all submitted tasks using this method. The participants came from a lot of different countries (varying from USA to India). Around 53% where from Europe with Spain having the highest share with almost 5%. Circa 19% of the workers where from Asia, 14% from India and 14% from USA. Figure 2 shows the final design of the task as presented to the workers.

# 5. DATASET DESCRIPTION

The dataset release includes 318 movie titles and links to gather online metadata and trailers for movies. We do not provide the video files because of copyright restrictions. The trailers were downloaded from IMDb [27] and YouTube [31]. Furthermore, we provide metadata collected from IMDb, Rotten Tomatoes [21] and Metacritic [13] including user comments.

The dataset includes also low-level visual, audio and text features extracted from trailers, posters, metadata and user comments. The provided visual features are Histogram of Oriented Gradients (HOG) gray, Color Moments, local binary patterns (LBP) and Gray Level Run Length Matrix [10]. The audio descriptors are Mel-Frequency Cepstral Coefficients (MFCC) [9]. For text information, we provide a term frequency–inverse document frequency ($td - idf$) matrix, which gives indications about the importance of different words [20].

The dataset enables evaluation of systems both with respect to the airline's choice of movies and the crowd's choice of flight-suitable movies. Votes about the labels collected by crowdsourcing are considered as the authoritative labels.
The development set contains 95 labeled movies. The test data contains 223 movies (the split is chosen based on what we think would provide a robust evaluation of algorithms tests with the dataset). Negative and positive classes in both splits of the dataset are balanced. The majority class baseline is 0.5 for precision. For the evaluation, we recommend standard metrics such as weighted average of precision, recall and weighted F1 score.

# 6. APPLICATION OF THE DATASET

To show the usefulness of the dataset, we conducted some initial experiments. The findings of these experiments are presented here. To confirm the viability of the dataset for supporting identification of movies suitable to be watched on an airplane, and show the possibilities that it opens we carried out some basic classification experiments. For these experiments, we used the WEKA machine learning library[2]. As a classifier, we choose the rule-based PART classifier.

This classifier uses separate and conquer to generate a decision list. From this, it builds a decision tree from which the best leaves are used as rules for the classifier [5]. Table 1 shows the results of our four initial experiments.

The first experiment (i) uses metadata (language, year published, genre, country, runtime and age rating) in combination with user ratings as input for the classifier. This run is our best performer. It clearly outperforms the naive baseline, which is 0.5 (precision, recall and F1-score).

The second run (ii), uses user ratings only (collected from IMDb, Rotten Tomatoes and Metacritic). This run performs well with recall, but poorly with precision. This implies that receiving certain user ratings is a necessary, but not a sufficient condition for being a movie that is good to watch on an airplane. Which is a very important message because it means, that using user ratings from standard platforms only does not lead to the best recommendations. This is a strong indicator that the watching situation is an important impact factor. Taken together, the first two runs confirm that the task is non-trivial, and that it is also viable.

The only user ratings experiment (ii) achieves virtually the same performance as Metadata + Visual run (iii). The results are not, however, exactly identical. We take this as motivation to perform further experiments in the future (different features, audio features, etc.).

The only visual information run (iv) uses global visual features for the classification. This run scores below the naive baseline. However, the approach to visual classification here was relatively simple. We only used one global image feature, namely Joint Composite Descriptor (JCD). JCD is a combination of Fuzzy Color and Texture Histogram (FCTH) and Color and Edge Directivity Descriptor (CEDD) [32]. It combines color, textural and edge information in one descriptor. This makes it a good choice for initial tests since the most promising parts are included. Additional exploratory experiments, not reported here, revealed that visual features do have the ability to approve results when used in combination with other features. Such combinations are interesting for future work.

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

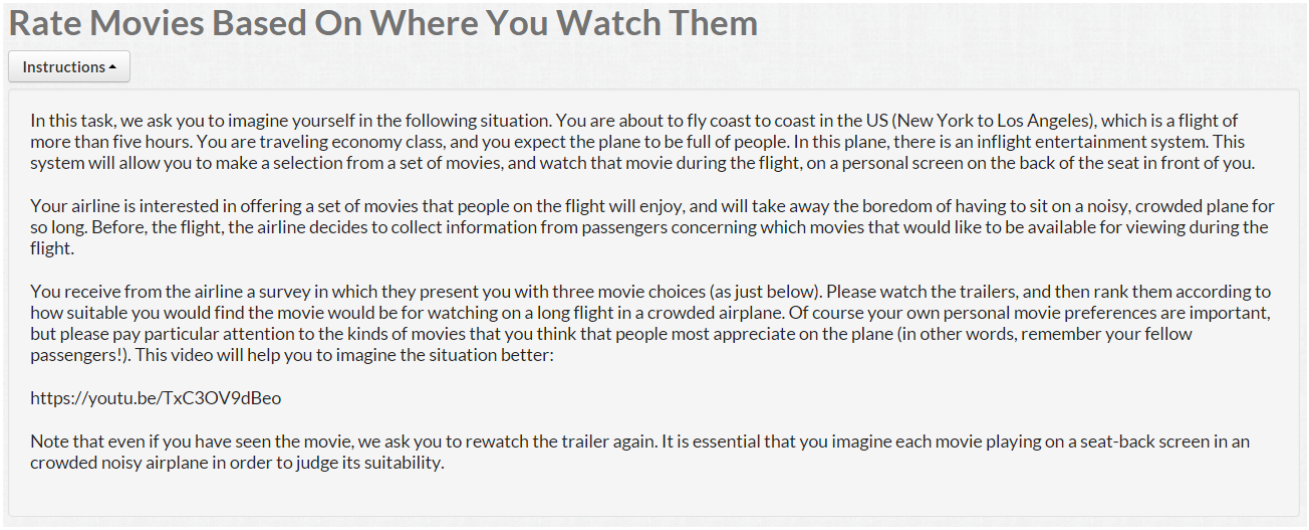



**Figure 3: Crowdsourcing task description. It also includes a link to a video that should help the workers to get in the feeling of a flight situation.**

Finally, the last experiment (v), using only metadata, confirms that metadata without user ratings is able to yield performance above the naive baseline. An information gain based analysis of all features ranked genre, publication year, country, language and runtime as the top five features.

## 7. LIMITATIONS OF THE DATASET

The collected data and the idea behind it is very novel and opens some promising directions in the field of multimedia. Nevertheless, it also comes with some limitations.

The crowd-sourcing study is carefully prepared with enough means to check for the subjects' reliability. However, the data for each movie are collected from five subjects only which can be seen as on the lower end considering the subjectivity and difficulty of the task. Moreover, this makes it hard for a statistical analysis which should be performed on any data collected from subjects.

Furthermore, the methodology of splitting the dataset into suitable and not suitable based on the ranks is questionable. To tackle this problem, all crowdworkers votes and rankings are included in the dataset. That should allow possible users a more detailed insight. Even though the task is well described for the observer, and the initial video to *place the subject into the situation* is well prepared, it is very hard to be sure that subjects fully understood the task and can picture themselves in the situation.

The data is also collected based on the trailers only, while the ranks from the databases are for the whole movies which can lead to some biases. A further limitation is that the data is only collected from one airline (KLM) so far. Although, investigating different airlines revealed that the used movies over the used time were almost identical.

This lead us to the conclusion that airlines most probably follow recommendations based on the popular rating sites. Taking all these limitation into consideration, we still believe that the obtained ground-truth data can give a first signal and open a new direction but any conclusions should therefore be drawn with taking them into account.

## 8. CONCLUSION AND OUTLOOK

We have presented *Right Inflight*, a dataset that allows researchers to explore the next challenge of predicting whether video content is suitable for a particular watching context. We choose to focus on airline travel, since the relative familiarity of the situation, and the relatively extremeness of the distractors, allow us to more easily tap into general opinions of people about the content suited for the context. The resulting dataset poses a challenge for multimodal classification that is extremely difficult. However, contrary to what one might expect, given the subjective nature of individuals' preferences for movies, inferring which movies are considered suitable for watching on an airplane is not impossible.

Our ambition is that the novel use case addressed by the dataset may inspire multimedia researchers to delve deeper into research questions that involve user viewing intent and the context of multimedia experience. As mentioned in the introduction, we believe that Context of Experience is important in helping people to decide which kinds of content is suitable for stressful situations including waiting rooms, airports, and during medical treatments, such as dental procedures. We hope that our dataset can help to raise awareness about the topic, but also provide an interesting and meaningful use case to researchers already working in related fields.

## 9. ACKNOWLEDGEMENTS

This work is partly funded by the FRINATEK project "EONS" (#231687) and the BIA project *PCIe* (#235530) funded by the Norwegian Research Council and by the EC FP7 project CrowdRec (#610594).

## 10. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.

[2] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. Liris-accede: A video database for affective content

analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

[3] A. Borowiak and U. Reiter. Long duration audiovisual content: Impact of content type and impairment appearance on user quality expectations over time. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 200–205. IEEE, 2013.

[4] Crowdflower. Crowdflower crowdsourcing platform. http://crowdflower.com. [last visited, March. 10, 2016].

[5] E. Frank and I. H. Witten. Generating accurate rule sets without global optimization. 1998.

[6] KLM-Royal-Dutch-Airlines. KLM on board entertainment system content. http://www.klm.com/. [last visited, March. 10, 2016].

[7] M. Larson, B. Ionescu, M. Sjöberg, X. Anguera, J. Poignant, M. Riegler, M. Eskevich, C. Hauff, R. Sutcliffe, G. J.F. Jones, Y.-H. Yang, M. Soleymani, and S. Papadopoulos. Proceedings of the MediaEval 2015 multimedia benchmark workshop. *The MediaEval 2015 Workshop*, 2015.

[8] P. Lebreton, A. Raake, M. Barkowsky, and P. Le Callet. Evaluating complex scales through subjective ranking. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pages 303–308. IEEE, 2014.

[9] B. Logan et al. Mel frequency cepstral coefficients for music modeling. In *Proc. of ISMIR*, 2000.

[10] M. Lux. Lire: Open source image retrieval in java. In *Proc. of MM*. ACM, 2013.

[11] M. Lux, D. Xhura, and A. Kopper. User intentions in digital photo production: A test data set. In *MultiMedia Modeling*, pages 172–182. Springer, 2014.

[12] MediaEval Benchmarking Initiative for Multimedia Evaluation. MediaEval benchmark homepage. http://www.multimediaeval.org/. [last visited, March. 10, 2016].

[13] Metacritic. Metacritics critics and ratings. http://www.metacritic.com/. [last visited, March. 10, 2016].

[14] B. Rainer and C. Timmerer. A quality of experience model for adaptive media playout. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pages 177–182. IEEE, 2014.

[15] J. A. Redi, Y. Zhu, H. de Ridder, and I. Heynderickx. How passive image viewers became active multimedia users. In *Visual Signal Quality Assessment*, pages 31–72. Springer, 2015.

[16] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank. Factors influencing quality of experience. In *Quality of Experience*, pages 55–72. Springer, 2014.

[17] M. Riegler, L. Calvet, A. Calvet, P. Halvorsen, and C. Griwodz. Exploitation of producer intent in relation to bandwidth and qoe for online video streaming services. In *Proceedings of the 25th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 7–12. ACM, 2015.

[18] M. Riegler, M. Larson, C. Spampinato, J. Markussen, P. Halvorsen, and C. Griwodz. Introduction to a task on context of experience: Recommending videos suiting a watching situation. In *Proceedings of the MediaEval 2015 Workshop*. CEUR-WS.org, 2015.

[19] M. Riegler, V. Reddy G, M. Larson, P. Halvorsen, and C. Griwodz. Crowdsourcing as self fulfilling prophecy: Influence of discarding workers in subjective assessment tasks. In *Proc. of CBMI*, 2016.

[20] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 2004.

[21] Rotten-Tomatoes. Rotten tomatoes critics and ratings. http://www.rottentomatoes.com/. [last visited, March. 10, 2016].

[22] A. Said, S. Berkovsky, and E. W. De Luca. Putting things in context: Challenge on context-aware movie recommendation. In *Proceedings of the Workshop on Context-Aware Movie Recommendation*, CAMRa '10, pages 2–6, New York, NY, USA, 2010. ACM.

[23] A. Said, S. Berkovsky, and E. W. De Luca. Group recommendation in context. In *Proceedings of the 2Nd Challenge on Context-Aware Movie Recommendation*, CAMRa '11, pages 2–4, New York, NY, USA, 2011. ACM.

[24] Y. Shi, M. Larson, and A. Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, 47(1):3:1–3:45, May 2014.

[25] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen. The MediaEval 2015 Affective Impact of Movies Task. In *Proceedings of the MediaEval 2015 Workshop, CEUR-WS.org*, 2015.

[26] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic. Corpus development for affective video indexing. *IEEE Transactions on Multimedia*, 16(4):1075–1089, 2014.

[27] The-Internet-Movie-Database-IMDB. Imdbs critics and ratings. http://www.imdb.com/. [last visited, March. 10, 2016].

[28] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[29] Tripinsurance. Best movies guide for airplanes. http://www.tripinsurance.com/tips/guide-to-the-best-moviestv-shows-to-watch-on-a-plane. [last visited, Dezember. 10, 2015].

[30] B. Winther, M. Riegler, L. Calvet, C. Griwodz, and P. Halvorsen. Why design matters: Crowdsourcing of complex tasks. In *Proc. of CrowdMM*. ACM, 2015.

[31] Youtube. Youtube movie sharing platform. http://www.youtube.com/. [last visited, March. 10, 2016].

[32] K. Zagoris, S. A. Chatzichristofis, N. Papamarkos, and Y. S. Boutalis. Automatic image annotation and retrieval using the joint composite descriptor. In *Informatics (PCI), 2010 14th Panhellenic Conference on*, pages 143–147. IEEE, 2010.

# Paper VIII

# Expert Driven Semi-Supervised Elucidation Tool for Medical Endoscopic Videos

# Expert Driven Semi-Supervised Elucidation Tool for Medical Endoscopic Videos

Zeno Albisser[1], Michael Riegler[1], Pål Halvorsen [1], Jiang Zhou[2],
Carsten Griwodz[1], Ilangko Balasingham[3], Cathal Gurrin[2]
[1]Media Performance Group, Simula Research Laboratory, Norway
[2]Insight, Dublin City University, Ireland
[3]Intervention Center Oslo University Hospital, University of Oslo, Norway
zenoa@ifi.uio.no, {michael, paalh, griff}@simula.no, {jiang.zhou, cgurrin}@dcu.ie, ilangkob@medisin.uio.no

## ABSTRACT

In this paper, we present a novel application for elucidating all kind of videos that require expert knowledge, e.g., sport videos, medical videos etc., focusing on endoscopic surgery and video capsule endoscopy. In the medical domain, the knowledge of experts for tagging and interpretation of videos is of high value. As a result of the stressful working environment of medical doctors, they often simply do not have time for extensive annotations. We therefore present a semi-supervised method to gather the annotations in a very easy and time saving way for the experts and we show how this information can be used later on.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## General Terms

Experimentation, Human Factors

## Keywords

video annotation, medical multimedia information systems, semi-supervised, object tracking

## 1. INTRODUCTION

Detecting irregularities in intestines is a difficult and very time-consuming task, and there are several different kinds of irregularities a doctor can detect visually using colonoscopy or camera pills. For the untrained eye, such irregularities are, however, not always easy recognizable. Depending on the length of the video acquired by, e.g., a camera pill, this can be a very time-consuming and therefore expensive task. It seems natural to try to automate this task using computers. To be able to train an algorithm to detect such irregularities, a comprehensive data set, containing video sequences with and without irregularities, is necessary. Collecting this data requires recording of video sequences and tagging every occurrence of an irregularity in every video frame, e.g., marking a polyp in the colon as shown in Fig. 1. This work requires a specialist to make sure that no false positives or false negatives examples occur. Tagging all the occurrences is an especially tedious piece of work, as it requires stepping through single video frames and adding, moving and resizing tags. The experts usually do not have a lot of time for this kind of work. Thus, makes it necessary to create tools that reduce the amount of time needed to process a video. Such tools must meet the following requirements: (i) Save as much of the specialist's time as possible, (ii) allow efficient collection of big amounts of data, (iii) easy to use with very little introduction time and (iv) deployment of the system in restricted hospital environment.

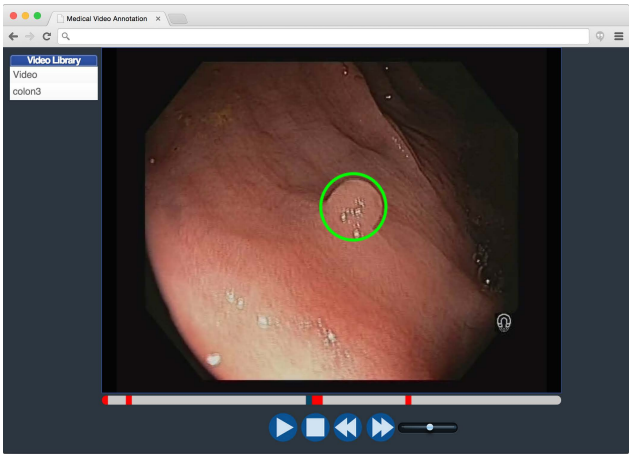



**Figure 1: Web based video annotation software for medical purpose.**

To tackle this problem we have been prototyping and experimenting with different technologies to cater these specific requirements. We present our semi-supervised annotation system, see Fig. 1, which is divided in two parts. The first part (i) is a web based tagging tool that should be used by a specialist to create a coarse selection of regions of interest. The second part (ii) is a tool that can be used subsequently by a regular user to generate a complete data set using object tracking algorithms and manual correction if necessary. The system is already in use for generating informative and large data sets for medical multimedia con-



*ACM MMSys '15,* Mar 18-20, 2015, Portland, OR, USA
ACM 978-1-4503-3351-1/15/03
http://dx.doi.org/10.1145/2713168.2713184.

tent analysis. The remainder of the paper gives an overview on related work and presents the architecture and the implementation of the application. Furthermore, we show how the obtained data can be used afterwards, and we give an outlook of ongoing and future work. A demo video of the tool can be found at `http://goo.gl/Fhd0J6`.

## 2. RELATED WORK

Previous research related to annotating videos can be split into manual video annotation tools and semi-supervised approaches. In this section, we will discuss their relation to our tool and point out the differences. A way of annotating videos is the use of different elements on top of the video frames like speech bubbles, hand drawn annotation and a lot of other different overlays. Furthermore, annotation by speech is also a widely used method. That these annotations have in common is that they are manually added to the video to describe the content. Examples for state of the art applications are, for instance, YouTube, VideoWiki and Popcorn Maker. A tool that combines complex annotations together is *Videojot*. For the medical use case, the *MedAnnotation Tool* is the latest related work in this area [14, 11, 4, 2, 13, 12]. The usage of these tools ranges from very complicated to very easy to use for, trained or untrained users. All these tools require a significant amount of time for creating annotations. In some areas, this is not a big problem, but in others like, the medical sector where the doctors are constantly under a lot of pressure and lack of time, the currently existing tools are not really usable, i.e., especially when the goal is to collect a huge amount of data for computer vision or retrieval algorithms [7].

Our tool tackles this problem by providing a very easy and quick way to annotate important parts. It then uses these tiny annotations to automatically generate the data that we need for further computation. There already exists some work about these kind of semi-supervised annotation tools, but they do not annotate specific parts of the video for the usage in a later training set. They are more general semantic annotation [5, 16, 15] tools, which cannot be used for example to detect cancer in regions of the video, etc. The biggest difference to existing tools is that the tool presented here is easy and time expeditious to handle, and it is able to automatically create a huge data set of medical conditions from a subset of expert annotations. Therefore, it supports the doctors to provide as much information as possible with very humble effort. To the best of our knowledge, there exist no such tool that provides the same functionality.

## 3. ARCHITECTURE

The architecture of our solution is divided into two steps *Manual Annotation* and *Object Tracking*. Fig. 2 gives an overview of the whole system. This is mainly to reduce the amount of time specialists are needed in the whole process due to the fact that they only have to provide elucidation in a single frame. We do require the specialist's knowledge during the first step to do a very basic identification of irregularities and to tag them accordingly. The *Manual Annotation* step is to precisely select any regions of interest in a video sequence. We also refer to this step as *Object Tagging*. The *Object Tracking* step is to track the regions of interest on previous and subsequent frames, based on the previously manually created tags. This step is more about

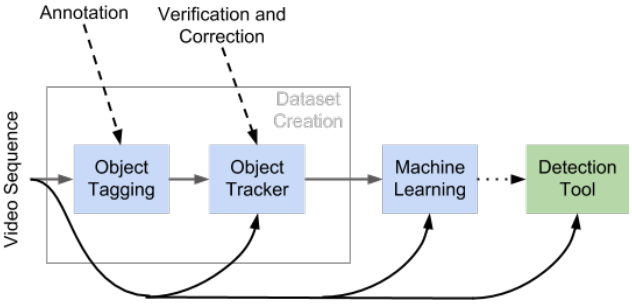



**Figure 2: The processes for dataset creation are a prerequisite for building and a detection tool.**

tracking an object and adjusting the size and position of the tracked region than about identifying or recognizing irregularities. Specialist's knowledge is therefore not required for the second step. Another reason to divide the process into these two steps is the technologies available for implementing the required software. A specialist is usually located in a hospital with special restrictions to security due to sensitive patient information. Deployment of software is therefore a problem because of privacy issues. Nevertheless, internet access and a browser are usually available. This makes standard web technologies a convenient way of circumventing deployment related issues for the manual annotation software. It also implies storing all information on the server side and moves the responsibility of maintaining the system and data integrity from the user to the server administrator.

**Manual Annotation.** The manual annotation is the first step in the whole data gathering process. In this step, a specialist uses rubber band selection (mark a bounded area) to create a coarse selection of regions of interest and annotates every selection with a name for classification. Every region needs to be marked once only. To keep the specialist's time spent on this task minimal, we do not require the region to be marked in the very first video frame it appears. Information on first appearance and change of shape or position within the picture will be added later using object tracking and manual correction. This approach allows a rather rapid way of working for the specialists. They might even watch the video at a higher playback speed and only stop or slow down the playback when really necessary. The information collected in this step includes the position and dimensions of irregularities in pixel coordinates, a classification and a timestamp relative to the beginning of the video for each selected region. We have decided to implement the manual annotation component using JavaScript and HTML5 video which is available in most recent web browsers. We use a standard username and password authentication mechanism and transfer all the date using HTTPS to ensure secure data access and transmission during the whole process.

**Object Tracking.** The output from *Manual Annotation* only contains a single tag for every region of interest in the video sequence. Using this information, we can now apply object tracking algorithms and manual correction to generate a complete data set. Most of the work in this step is done by the software. The user just needs to step 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. As the irregularity most likely has not been marked on the very first frame it appears in, the video must also be played in reverse direction from the first position a region was marked. This is needed to track the region towards the beginning of the video. There is of course still a fair amount of manual work involved in this task. However, using a suitable tracking algorithm, the time needed to create a complete dataset can be reduced significantly. Moreover, specialist skills are usually no longer required here as the whole task is simply about tracking regions and adjusting rectangular dimensions rather than actually detecting or recognizing irregularities. The output generated in this step is a list of rectangles for every previously marked region. Every rectangle in such a list is described by the index of the video frame it belongs to, its position in pixel coordinates and its dimensions.

# 4. IMPLEMENTATION

Experimenting with several different technologies we came to the conclusion that a solution divided in two steps has several advantages. It allows us to minimize the time a specialist is needed, and it also significantly simplifies the deployment and maintenance of the software. The only requirements for the first step are an HTML5 compliant web browser and an internet connection.

**Manual Annotation.** The web application we implemented is mostly written in HTML5 and JavaScript. Specifically, it makes use of the HTML5 video element. Listing and uploading videos and storing tagging information is implemented in Java and running in an Apache Tomcat servlet container[1]. All video sequences will be uploaded to the server through the web interface. On the server, we are using a Java servlet, which spins off a job to transcode the video to H.264. For transcoding we are using *libav* and *avconv*[2]. Transcoding is necessary in case the original video file is not encoded in a codec that is supported by the browser. H.264 seems to be a good choice as it is currently supported by all major web browsers. The transcoding job is running asynchronously, so a connection to the server is not needed to keep the job alive.

The web interface of our tagging application provides the usual start, stop and pause controls of a regular video player. Additionally, we added a seek bar that highlights the playback position and any regions of interest in colors. We also added a "seek-forward" and a "seek-backward" button that allows stepping to the next/previous region of interest. As the video playback in HTML5 is running outside of the JavaScript execution thread, we do not have a strict control over the video frames being displayed. The playback position is only provided as a floating point value property *currentTime* in seconds. The property can be read and it can also be written in order to seek to a specific position. When executing JavaScript code this property can be read at an arbitrary point in time. And since a single video frame is usually being displayed for about 40ms [3] this means that when playing a previously tagged video sequence, we will most likely not read the same value from the *currentTime* property again as we were reading while tagging. Therefore visibility of a previously created tag cannot be guaranteed

[1] `http://tomcat.apache.org`
[2] `https://libav.org/avconv.html`
[3] assuming a usual frame rate of 25 frames per second

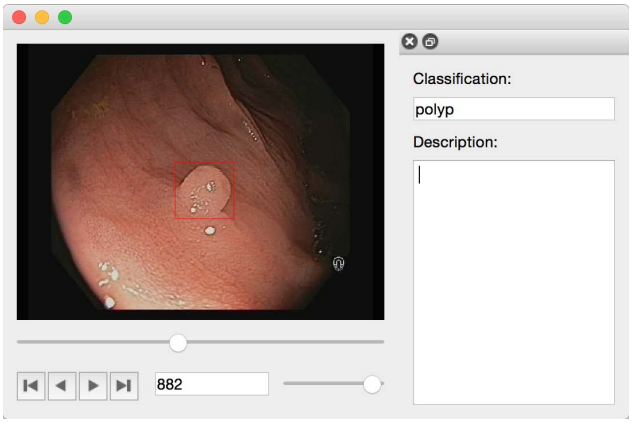



**Figure 3: Native software for modifying tags and tracking of regions of interest.**

during playback and we must use the seek buttons to seek to the next or previous region of interest. Whenever a region of interest has been selected, an editor shows up and allows the specialist to enter a classification and a comment. This information will be stored together with the tagged rectangle in JSON[4] format on the server.

**Object Tracking.** For the second step in the process, we implemented the object tracking tool in C++ using Qt[5] for the user interface and OpenCV[6] for reading and processing the video data. We further integrated with Struck [8] for tracking the tagged regions. The user interface for this tracking software is similar to the web interface described previously and can be seen in Fig. 3. It features a video widget, play, seek-forward and backward buttons as well as a seek bar with identical behavior. Moreover, a slider to increase or decrease the playback speed and an editor for classification and description is present. Further, a button for playing the video in reverse direction and context menus for modifying regions of interest is provided. After starting the application, a JSON file created using the tagging web application can be opened and the respective video file must be selected. We use the original video file instead of the H.264 encoded one. This is because we need to be able to play the video forwards and backwards frame by frame. Recreating frames in reversed direction is very expensive with H.264, because frames can be encoded referencing previously encoded ones. The original files uploaded to our server are usually simple MJPEG video files and are very well suited for playing both directions. The users use the seek buttons to seek to the next or previous regions of interest. Then they use the context menu to select one or multiple regions for tracking. Playing the video in either direction will then track the region in the video frames being displayed. Alternatively, the arrow keys can be used to step forward or backward frame by frame. The playback can be paused at any time to adjust size or position of the tracked region.

Using double buffering allows reading and processing the next frame while the previous frame is still being displayed. The processing (reading of frames and tracking of regions) is therefore running in a separate thread. The communication between the user interface and the worker thread is imple-

[4] `http://goo.gl/Oi5kIF`
[5] `http://www.qt.io`
[6] `http://www.opencv.org`

mented using Qt's events delivery mechanism. Whenever the tracking algorithm fails to track a region, the playback stops automatically. It is then up to the user to decide if the tracked region should be removed or if the tracking should be re-initialized with an updated region. The user can seek forwards and backwards freely to review the tagging and tracking results and adjust, move or restart tracking of a region at any point during the process. Once the dataset is complete, it can be saved to a JSON file.

# 5. APPLICATIONS OF THE DATASET

The primary application of the annotated images is training algorithms for automatic medical screening. As stated at the begin, reviewing images or videos and making diagnostic decisions in screening are very time-consuming and the accuracy is subject to the experience and concentration of the physicians [6]. For example, in a camera pill endoscopy exam, there are about $60,000$ images per examination for one patient, and it costs an experienced medical clinician about 2 hours on average to view and analyse all the video data [10]. Therefore, it becomes necessary to reduce the heavy burden on physicians and speed up the screening process with computer aided diagnosis. In terms of colonoscopy videos, the objective would be training a classifier and automatically detecting the colon cancer, or its precursor lesions, colorectal polyps in videos. To build the classifier, the annotated irregularity regions are pooled together as positive samples and random selected regions without any irregularity are used as negative samples. Colour, texture and shape features [1, 3] are extracted from the training samples. A Support Vector Machine (SVM) is used to train the classifier with the combinatorial features, and the Radial Basis Function is applied as the kernel [9]. To tune the parameters in SVM and prevent model over-fitting, k-fold cross validation is performed. A separated set of positive and negative samples, which have never been seen during the training, is prepared as a testing set. The classification performance is then measured by the Receiving Operating Characteristic (ROC) curve. With a shifting-window method, the built classifier can not only tell the presence of irregularities but also give their locations within an image. Beside the automatic screening, with our semi-supervised annotation tool, segments within a medical video are marked and labeled with specialists' knowledge input. Such annotated videos can be directly used in medical video archive for surgical documentation.

# 6. CONCLUSION AND FUTURE WORK

In this paper, we presented an application for annotation of any kind of videos that need expert knowledge for the elucidation. We focused on the medical use case of endoscopic videos. The time that doctors have to spend with this tool to annotate the videos is extremely low. Furthermore, we showed that the tool is able to automatically create more annotations based on the initial annotation by the experts and how these annotations can be used. It provides a possibility for easy annotation for further analysis, documentation or lecturing. In the future, we will focus on gathering a large dataset and the usage of it in machine learning or computer vision algorithms. We further would like to expand the use case to other domains like sport.

# 7. ACKNOWLEDGMENT

This work is founded by the FRINATEK project "EONS" (#231687) and the iAD Center for Research-based Innovation (#174867) by the Norwegian Research Council.

# 8. REFERENCES

[1] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino. Texture-based polyp detection in colonoscopy. pages 346–350, 2009.

[2] W. Bailer and P. Schallauer. Detailed audiovisual profile: enabling interoperability between mpeg-7 based systems. In *Proc. of MMM '06*, 2006.

[3] M. K. Bashar, K. Mori, Y. Suenaga, T. Kitasaka, and Y. Mekada. Detecting informative frames from wireless capsule endoscopic video using color and texture features. In *Proc. of MICCAI'08*, pages 603–610, Berlin, Heidelberg, 2008. Springer-Verlag.

[4] M. Bastan, H. Cam, U. Gudukbay, and O. Ulusoy. Bilvideo-7: an mpeg-7-compatible video indexing and retrieval system. *MM, IEEE*, 17(3):62–73, 2010.

[5] M. Bertini, A. Del Bimbo, and C. Torniai. Automatic video annotation using ontologies extended with visual information. In *Proc. of ACM MM'05*, pages 395–398. ACM, 2005.

[6] B. Giritharan, X. Yuan, J. Liu, B. Buckles, J. Oh, and S. J. Tang. Bleeding detection from capsule endoscopy videos. In *Proc. of EMBS'08*, 2008.

[7] M. Guugenberger, M. Riegler, M. Lux, and H. Paal. Event understanding in endoscopic surgery videos. In *Proc. of ACM HuEvent'14*. ACM, 2014.

[8] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In *Proc. of ICCV'11*, 2011.

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

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

[11] C.-Y. Lin, B. L. Tseng, and J. R. Smith. Videoannex: Ibm mpeg-7 annotation tool for multimedia indexing and concept learning. In *ICME '03*, pages 1–2, 2003.

[12] M. Lux and M. Riegler. Annotation of endoscopic videos on mobile devices: a bottom-up approach. In *Proc. of ACM MMSys'13*, pages 141–145. ACM, 2013.

[13] M. Riegler, M. Lux, V. Charvillat, A. Carlier, R. Vliegendhart, and M. Larson. Videojot: A multifunctional video annotation tool. In *Proc. of ACM ICMR'14*, page 534. ACM, 2014.

[14] R. Schroeter, J. Hunter, J. Guerin, I. Khan, and M. Henderson. A synchronous multimedia annotation system for secure collaboratories. In *Proc. of e-Science'06*, pages 41–41. IEEE, 2006.

[15] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proc. of CVPR '10*, pages 966–973. IEEE, 2010.

[16] T. Zhang, C. Xu, G. Zhu, S. Liu, and H. Lu. A generic framework for video annotation via semi-supervised learning. *MM, IEEE*, 14(4):1206–1219, 2012.

# Paper IX

# Event Understanding in Endoscopic Surgery Videos

# Event Understanding in Endoscopic Surgery Videos

Mario Guggenberger
mg@itec.aau.at

Mathias Lux
mlux@itec.aau.at

Institute of Information Technology
Alpen-Adria-Universität Klagenfurt
9020 Klagenfurt am Wörthersee, Austria

Michael Riegler
michael@simula.no

Pål Halvorsen
paalh@simula.no

Media Performance Group
Simula Research Laboratory AS
Norway

## ABSTRACT

Event detection and understanding is an important area in computer science and especially multimedia. The term event is very broad, and we want to propose a novel event based view on endoscopic surgeries. Thus, with the novel view on surgery in this paper, we want to provide a better understanding and possible way of segmentation of the whole event *surgery* but also the included sub-events. To achieve this sophisticated goal, we present an annotation tool in combination with a thinking aloud test with an experienced surgeon.

## Categories and Subject Descriptors

H.5 [**Information Interfaces and Presentation (e.g., HCI**]: [Miscellaneous]

## Keywords

annotations; events; video; endoscopy; event understanding

## 1. INTRODUCTION

Understanding events can leverage the development of automatic algorithms for learning, detection, or classification to a high degree. When hearing the word *event*, people usually think of high-level events like concerts and parties, but even a surgery procedure on the heart can be seen as an event. This obviously leads to the conclusion that events are hidden everywhere. In this paper, we take an event-based look at endoscopic surgeries, or more specifically, the annotation of videos of laparoscopic surgeries. Our findings should, however, be applicable to different types of endoscopic interventions.

Endoscopic surgeries can be seen as a special type of human centred event since they involve the participation of multiple people. They are very complex, and a lot of expertise is necessary to annotate recordings of the operations, which is the reason why surgeons usually do it themselves. The rationale to annotate events varies, but it is usually for documentation or training purposes. The problem is that the currently provided tools are very often too complex for surgeons and make it hard to capture the important information in a fast way, or there are no tools provided at all. Since surgeons are often under a lot of time pressure, they want to annotate their surgeries in an intuitive way and as fast as possible. Therefore, it is important to assess their requirements as soon as possible and include their expertise in the design phase to decide which functions and results are important and would support them later in production use. Fancy features, like instrument detection, do not help much if they do not provide useful information to the surgeons and doctors that will finally work with the results.

To tackle these problems, we designed an annotation tool that is able to support the doctors to annotate surgeries naturally. We then performed a thinking aloud test with a

**Figure 1: Experimental setup of the thinking aloud test.**


*HuEvent'14,* November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2660505.2660509.

world renowned surgeon working in the field of laparoscopy to obtain first-hand information about what such an annotation tool needs to provide in order to be suitable for and usable by surgeons and lead to a better understanding. The thinking aloud test is a proven method to test user interaction with a system. It requires a special setup including the recording of the interaction with the program and the reaction of the person itself. Figure 1 shows this experimental setup. One camera was used to record the doctors reactions (small image in the upper right corner) and one to record his interaction with the annotation tool (main image). Our work shows that different kinds of events in surgeries ask for different kinds of annotations, distinguished by their level of detail. We also present an event-based segmentation model of endoscopic surgeries, based on the analysis of our expert's information. We believe that our work will help researchers to collaborate and get information from surgeries more efficiently. The main contributions of our work therefore are:

- Providing a general event-based model that is valid for different types of endoscopic surgeries.
- Providing detailed information about which functions an annotation tool for surgeries should include.
- Presenting an advanced prototype of the annotation tool.
- Providing a better understanding of endoscopic surgeries with the help of an expert.

Note at this point that the annotation and understanding of the surgeries is just a first step. The collected information will be used for machine learning and computer vision techniques, and in the best case lead to automatic detection or classification of events or sub-events.

In the reminder of the paper we will at first give an overview and discuss related work in the area of events and endoscopic surgeries. Then we present the methodology split into an annotation tool and a thinking aloud test. In the evaluation section we present the analysis and findings of the experiment in a conceptual and technical point of view. Finally we draw a conclusion and discuss about ongoing and possible future work based on our insights of this paper.

## 2. RELATED WORK

Philosophy defines an event as a special incident over a specific time span involving one or more objects and happening at a specific place, which can be described by an encasing term - the name of the event. For example, there are general events like endoscopic surgeries, birthdays or funerals, and sports events like football, basketball or soccer games. An event can also consist of many sub-events. For instance, a soccer game has goals and fouls as sub-events and an endoscopic surgery consists of sub-events like injections or cuts. Moreover, an image may depict an event, but it is usually just a snapshot and therefore only covers one time instant of the event's time span [8].

In the context of event processing, detection is an important field of research. It is widely employed in computer vision and classification, because to classify an event, it needs to be detected before. Reuter et al. [13] are using the concept of events to classify multimedia streams automatically into corresponding events. To achieve this, they use a two-step approach. At first, they retrieve event candidates, and secondly, they use machine learning to assign new event candidates to existing ones or to new ones. Petkos et al. [12] try to tackle social event detection by presenting an algorithm that uses multimodal clustering and multimodal fusion to combine different features that can be helpful for event detection in a clever way. A similar approach is presented in [15] by Zeppelzauer et al. They use an unsupervised clustering method to cluster events based on time, user and geo-location information. All these approaches are well performing state-of-the-art methods, and they show that event-based segmentation or classification are promising directions for multimedia content exploring.

Another important direction in the research field of events is event synchronization. Event synchronization combines data of different sources to form an overall picture of a specific event, which usually includes pictures and videos. This can help to get a better understanding of events. Actual work form this area is presented in [4] and [7]. In the first paper the authors try to analyze the content of the images in different photo collections to synchronize them into homogeneous events. The second paper describes an approach to synchronize streams of photos based on to which events they are belong. This is done by a scalable message-passing based optimization framework. Additionally, there are initiatives like the MediaEval Benchmark[1] with tasks like social event detection and multi-user event media synchronization, which shows that the consideration of events is a promising and interesting field of research.

To the best of our knowledge, there is no work that approaches endoscopic videos as a flow of events like we propose in this paper. In [11], Münzer et al. take a low-level bottom-up approach by detecting three classes of irrelevant segments in endoscopic videos (dark, out-of-patient, blurry). Transitions between those classes and the *relevant* class can be seen as low-level sub-events, e.g., the start of the actual surgery when the first out-of-patient segment transitions to an in-patient segment. In this work, we facilitate the detection of high-level sub-events in surgeries through annotations by the actual surgeon. As stated before, regarding a surgery as an event with hierarchical sub-events can help to understand the surgery better, and it can provide an easy view on a complicated topic that is understandable by both surgeons and computer scientists. We anticipate that this increased understanding can lead to the development of better annotation tools, which in turn can provide better information for computer vision, machine learning and classification approaches.

Overall, there have not been many attempts to use multimedia content like images or video for a better understanding of events in the medical sector. Battles et al. [1] presented an event reporting system for blood transfusions. Their system was designed to detect, select, describe, classify, compute, interpret and locally evaluate the event of blood transfusion. It was shown that such a system can improve the health-care results positively, but a good system strongly needs input from both end-users and external experts. The reason is that doctors often have their own techniques of handling multimedia material in their hospitals, which is hardly ever the most sophisticated or effective way. Due to a lack of knowledge in computer science, they often underestimate or do not know the capabilities of

[1] `http://www.multimediaeval.org/mediaeval2014/`

techniques like machine learning and computer vision. An interdisciplinary working expert with extensive knowledge in the area is therefore desired as a source of information, ideas and ready-to-use techniques.

Since endoscopy covers a large number of different surgeries like gastroscopy, mediastinoscopy, rhinoscopy, colonoscopy, laparoscopy, and arthroscopy, it is still a rather unexplored field in multimedia. The most explored sub-type are colonoscopic surgeries. The latest state of the art for colonoscopy images and videos is 3D reconstruction of the colon like discussed in [5, 6]. Current related work regarding videos of endoscopic surgeries in general can be found in [2] where the authors automatically segment a surgery into distinct phases. Furthermore, Münzer et al. [10] are concerned with the detection of the typical circle that is framing the view of an endoscope. We expand this knowledge with a description of the general procedure of an endoscopic surgery event and its encompassing sub-events.

# 3. METHODOLOGY

The methodology that we were employing was a thinking aloud test setup as described in [3]. It consisted of two stages, stage one being a hands-on experience by a surgeon who used our annotation tool as he would wish to use it in his daily work. Stage two was an open interview reflecting his experience with the tool, and an interview following a prepared exit questionnaire where we asked specific questions that came up during the creation of the concept and the implementation of the tool.

We did this test with only one expert because it is a highly specialized domain where experts are scarce resources and hospital doctors in general have very limited time. Our expert from a regional hospital is a lead technology user in this area who has been recording, documenting, and even live-broadcasting his surgeries overseas since many years. Due to storage demands, he does not always record the full coverage of a surgery (the 1080p format with constant bit rate as recorded by the employed equipment has a high storage demand), but sometimes limits the recordings to cover only the most important phases of a surgery. To mark important moments, he additionally saves single frames as pictures. Both pictures and videos are not only used for the hospital's internal documentation, but also to explain surgeries to patients, to present and discuss interesting cases with colleagues and at conferences, and to school student trainees. Currently, he does not have any means to annotate his recordings and to store them. Persisted annotations would not only save him time to repeat an explanation, but make it more tangible to presentees, enable iterative improvements of annotations, and automatically build up a library of annotated videos. Such a library will then help computer scientists to analyze and classify events, detect similar events in unannotated videos, segment videos any hopefully even semantically synchronize them.

## 3.1 Annotation Tool

The annotation tool as seen in Figure 2 is an improved version of the annotation tool described in [9]. The most important requirement of the tool was to make its usage as simple as possible, and at the same time, extract as much information as possible from the video. The tool is a tablet computer with a video player that loads a recording of a surgery, and offers four main functions: (i) drawing visual annotations by hand or by using a touchscreen pen, (ii) recording spoken audio notes from the microphone, (iii) setting bookmarks for easier navigation in the video, and (iv) providing a video timeline that visualizes existing annotations and also enables navigation inside the video.

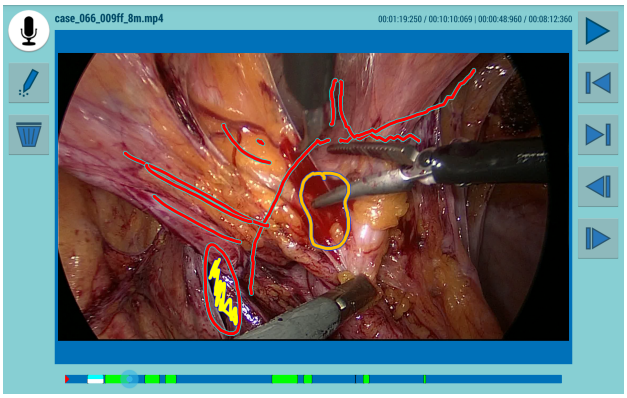



**Figure 2: The interface of the annotation tool including annotations of the doctor. The buttons on the right side are for controlling the video, like seeking forward or backward, or switching between existing annotations. The buttons on the left side are used to make annotations or delete them. By pressing and holding the annotation button (pencil), the doctor can start the annotation. While the button is pressed, the annotator can draw annotations in the video and the voice is recorded. The annotation can be stopped by releasing the button. This can be done while the video is playing (moving annotations) and while the video is paused (still annotations). The timeline at the bottom can be used for seeking and also shows already existing annotations and their duration.**

Hand drawn and spoken annotations are supported for video playback (moving annotations) and for still frames when the video is paused (still annotations). Annotations on still frames extend the total runtime of the video when played back because the annotations are animated in the same way as the surgeon did draw them and the spoken annotation is replayed correspondingly. Technically, both hand drawn and spoken annotations are recorded in parallel, i.e., it is not possible to explicitly select one or the other, but the annotator can still freely decide his means to annotate. For each stroke of the hand drawn annotations, the color can be selected from a predefined color palette. Bookmarks can be set by shaking the tablet quickly to mark important frames in the video. When playing back the video with its annotations, the hand drawn strokes are animated like they were drawn, and the audio is overlaid. We also have an HTML5 player capable of playing back the annotated video on the web, which has been extended for a different use case [14].

## 3.2 Hands-On

For the hands-on experience, we handed a Nexus 7 tablet running our annotation tool to the surgeon. We asked him to use it and annotate the videos as he envisions, and to speak out his thoughts while doing this. Since we had ac-

cess to the recordings of his surgeries, we randomly selected three videos recorded during the past 12 months, out of a collection of hundreds of recordings. Two of them are clips of 8 and 12 minutes runtime and the third is a full surgery recording with a runtime of about 2 hours. We are sure that the thinking aloud protocol, where the user assesses the tool while using it, is much more effective than a usual expert interview because it exposed problems that nobody of us had thought of yet, and it enabled a qualitative investigation of the tool and its annotations. The session was recorded with two video cameras, one over his shoulder capturing his mechanical interaction with the tool, the other from the front, capturing the whole scene including his face and voice.

## 3.3 Interview

The interview was also recorded by the same cameras and started with a discussion of the hands-on experience. It gave us a lot of insight on the expert's expectations of such a tool and the chance to discuss possible solutions to arisen problems. We then concluded it with a prepared exit questionnaire, where we tried to assess his satisfaction with the tool, possible usability features, use-cases, video processing methods, navigation patterns, and its market or *everyday use* potential.

# 4. EVALUATION

The video recordings of our thinking aloud session have been investigated by a group of three people to learn the most of the session. While an evaluation with a single expert cannot be fully valid for all people in this medical domain, it was still very productive and definitely showed us a precise direction we need to take. We define a model that is a valuable base for a study of larger scale. We divide our insights into conceptual findings that apply to the whole area of surgical event annotation, and technical findings that apply to the implementation of our tool, but which can still be valuable to developers of similar tools.

## 4.1 Conceptual Findings

A totally unexpected, but perhaps the most interesting insight that we got from our evaluation is the idea of a general event model of endoscopic surgeries, where the granularity of the events is directly connected to the type of annotation. We observed that our test candidate followed a pattern on all videos, where he always annotated the same kind of event with the same kind of annotation. The model is shown in Figure 3. A surgery can be split into different hierarchical sub-events. The first two sub-events can help to segment an operation in-patient and out-of-patient. If the camera is outside the patient, the segment is not interesting and does not carry any medical information. The surgeon did never perform an annotation when the camera was outside the patient. Therefore, it makes sense to segment a video based on that first.

When the camera is inside the body of the patient, there are three possible sub-events. A surgery is usually started with an overview of the concerning area, to document the actual status of the body and objects of interest, followed by the actual surgery. It is concluded by another overview after the surgery is finished, which leads to a documented before-and-after comparison.

During the surgery, we can first distinguish between general actions. These are moving around, which leads to blurry

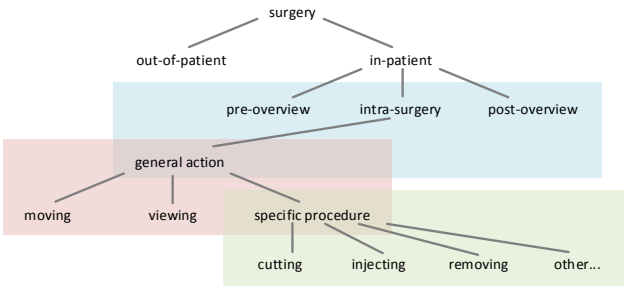



**Figure 3: Model of the endoscopic surgery event. The higher the event in the hierarchy, the lower is the level of detail and granularity required for the annotations. The blue box marks low detail events (interval-marking annotations), the red box marks medium detail events (speech annotations to running video), and the green box marks high detail events (speech and hand drawn annotations to still frames).**

and useless material, and viewing around. Moving around is for example when the doctor moves the camera in the body or through the colon to a specific area. Viewing around is when the surgeon looks around in a specific part of the patient like in the stomach or in an area of the colon where for example a disease is located. For these two events, the surgeon used only the annotation by speech. It could make sense to differentiate between these two types, in order that moving-around or looking-around parts in the videos can for example automatically be skipped, stored in a lower resolution or bit rate, or replayed with higher speed. The highest-granularity event that we identified during a surgery is a specific medical procedure. This can be an action like cutting a cyst, injecting liquid, removing a polyp, but also detecting exceptionally normal or abnormal looking organs (e.g., a liver without any disease and one with cancer). When one of these events occurred, the surgeon always paused the video and described it with both hand-drawn annotations and speech.

These findings clearly show that a division of the surgery into sub-events makes sense. For hierarchically high events, annotations do not have to be very detailed, but even imprecise descriptions can provide important information for classification or segmentation algorithms that have the potential of helping surgeons. In contrast, specific sub-events like cutting and injecting require very detailed annotations. This provides three very important advantages, namely the surgeon's usage for teaching purposes, the potential of automatically generated summaries, and the researchers' usage for training specific algorithms like cancer detection or instrument detection.

Our tested tool offered just one kind of annotation which could be universally applied in every situation: recorded speech combined with hand drawings, which are both recorded but optional, that ultimately covers a segment of the video. Since it seemed cognitively demanding of our test candidate to decide which annotation to take, we propose to shift to an event-based paradigm offering three kinds of annotations that cover three levels of detail: (i) marking intervals as the lowest detail level, (ii) recording speech to the running video as medium detail, and (iii) recording speech

with hand-drawn schemes on still frames for the highest detail level.

As seen in [11], low-level annotations can already be generated automatically, and we suggest it as a preprocessing step, of which the resulting intervals should automatically be integrated as annotations into the tool's timeline. Additionally, it is planned that interesting events will already be marked during the surgery, which makes them much easier to be found and annotated afterwards. Also interesting is the fact that the surgeon is not interested in classical image-processing methods like measuring the sharpness of the picture, the intensity of movement, dominant colors or surgical instrument detection. The only feature he would be interested in is the out-of-patient detection. He explained that recording this kind of surgeries works analogue to a movie script: the best and most interesting shots are usually deliberately orchestrated, meaning a stable camera, good lighting and no instruments blocking sight.

Finally we want to point out that, for the doctor, annotations on a playing video were useless in the current version (i.e., the surgeon always paused the video before starting a new annotation). They would be a very important feature if the annotation system would support the surgeon during the annotation process with object tracking that automatically repositions the drawings according to the camera movements. Otherwise he would have to do a nearly frame-by-frame-wise correction of the annotations, which would be too tedious.

## 4.2 Technical Findings

On the technical side, a big issue was the tablet size of 7 inches, which is great for portability reasons but a trade-off between user interface widgets and the video, i.e., the annotation drawing area size. Our tester indicated that a tablet with at least a 10 inch screen would be preferable. An additional improvement can be achieved by elaborating a usability concept where the video is drawn over the entire screen and control elements overlaid when needed, like it is typical for video players in full screen mode.

The next major issue is the drawing of annotations by fingers. The first thing our test candidate asked for when beginning the test was a touchscreen pen. This might be a question of individual taste, but this candidate definitely had problems drawing with his fingers. On the small tablet, a pen has the advantage that it does not occlude as much of the screen as fingers do. We discussed the usage of proprietary technology like the Samsung S-Pen[2], which would enable advanced drawing techniques like thickness adjustments of drawn lines by pressure. However, at the same time, it takes away the tablet screen's multi-touch ability that our tool is currently designed for.

The biggest software issue we encountered during our test was the video player provided by the Android 4.3 API, which does not support seeking to exact frames, but rather to the nearest sync frame. During our internal tests, we did not notice this issue as our video's group of pictures size was small, and errors of a few frames did not stand out. The surgeon, however, noticed even misplacements of single frames which, according to him, were a great distraction making his precise annotations worthless.

Regarding general usability, care must be taken that surgeons are usually not that computer savvy and do not have a lot of experience with different apps and platforms, which means that they are not used to common user interface paradigms. As an example, indicators of the recording status in the Android action bar, or Android toast messages as action feedback, often went unnoticed by our test candidate. This needs to be addressed with new concepts of greater visual and maybe even tactile impact. The video timeline as shown before is also not intuitively useable with long running videos, as annotation markers tend to shrink too small and get packed together. The solution might be a separate zoomed section of the timeline around the current position, to preserve a good level of overview detail.

To draw annotations, our tool offered different colors for one single type of stroke. It turned out that our tester used the colors very sparingly and only changed them randomly for no conscious reason, as he told us. We observed though, that the types of usage of the stroke can be classified into three different actions: (1) marking borders with a solid stroke, (2) marking areas with dots or hatched lines, optionally surrounded by a solid stroke, and (3) indicating directions of actions by drawing arrows from solid strokes. These actions always directly relate to the spoken annotation. We figured that it would be more helpful to provide different drawing tools with a single color each, instead of one tool with multiple colors. These could be a thin pen, a thicker felt pen, and a very thick semitransparent marker to highlight areas.

We also discussed the possibility to completely separate voice from drawn annotations, since they are now intertwined and, e.g., deleting an annotation deletes both the audio and the strokes, which is not always desired. This would however lead to two different annotation timelines that both want to be mapped to the video timeline and bring up many open questions. What if the annotator records his voice to the running video, then in parallel pauses the video for a drawn still-frame annotation, and later deletes this drawn annotation? Should the tool display a paused frame for no obvious reason, should this interval be cut out from the voice track to retain synchrony with the remainder of the voice-annotated video, or can we afford to lose synchronization between the voice annotation and the running video when purging the paused interval? There are several other cases leading to such situations.

There are also several potential additional features that our candidate indicated as helpful. Zooming into the video is anticipated since the interesting action often happens in a limited area in the video, and the assistant filming the surgery does not always correctly zoom in. This feature would go without a lot of image quality loss on the relatively small tablet screen as the videos are usually recorded in 1080p format. It would help focusing on the important area, drawing annotations more precise and also generate interesting metadata for analysis. He also mentioned the possibility to export and share annotated still frames from the videos, to fast-forward unimportant segments, and play back important ones in slow motion. He also wished for a function to render the annotations into a standalone video file. Lastly, he envisioned a multimedia integration of external image material from x-ray, ultrasonic and magnetic resonance therapy, to use the tool for multimedia presentations. He also thinks about usages for measurements in standardized recording settings, e.g., measuring size and area or analyzing structure and color of liquids and tissue.

[2] `http://developer.samsung.com/s-pen-sdk`

## 5. CONCLUSION

We presented an evaluation of a tool to visually and vocally annotate videos from endoscopic surgeries, evaluated with the help of a high-class surgeon in the field and on his own recordings. We deduced a general event model of such a surgery and identified a direct relationship between the granularity of an event and the type of its annotation. We also provided many insights, ideas, and a better understanding of endoscopic surgeries. They can help to develop appropriate annotation tools, which then in turn yield several interesting data and metadata for the analysis, classification, and post-production of endoscopic videos.

Regarding future work, we want to iteratively develop our tool to a state where it will be productively used by at least our collaborating surgeon but hopefully by his colleagues as well. Through this, we hope to collect a huge pool of surgical event annotations that we want to analyze and hope to use for the automatic detection of similar events, and ultimately for the retrieval and the semantic synchronization of similar video recordings.

Furthermore, we want to test more sophisticated approaches in combination with the annotation tool. For example, we are testing annotations supported by object tracking in real time and a frame-by-frame based annotation where the video can be slowed down. We develop these two approaches based on web technologies. They are currently in an experimental stage, and we are discussing the applicability with several surgeons at a large hospital. Once they are ready for testing we also would like to perform a parallel thinking aloud test with a larger number of doctors.

Our findings will help to improve both the quality of our annotation tool and the data generated from it, and we are sure they will help other researchers working in this area.

## 6. ACKNOWLEDGEMENTS

We would like to thank Prim. Univ.-Prof. Dr. Jörg Keckstein from the Villach Regional Hospital for taking the time to work with us. This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria, and funding from the European Regional Development Fund (ERDF) and the Carinthian Economic Promotion Fund (KWF) under grant 20214/22573/33955 and by the iAD center for Research-based Innovation (project number 174867) funded by the Norwegian Research Council.

## 7. REFERENCES

[1] J. B. Battles, H. Kaplan, T. Van der Schaaf, and C. Shea. The attributes of medical event-reporting systems. *Arch Pathol Lab Med*, 122(3):132–8, 1998.

[2] T. Blum, H. Feußner, and N. Navab. Modeling and segmentation of surgical workflow from laparoscopic video. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 400–407. Springer, 2010.

[3] T. Boren and J. Ramey. Thinking aloud: reconciling theory and practice. *Professional Communication, IEEE Transactions on*, 43(3):261–278, Sep 2000.

[4] M. Broilo, G. Boato, and F. G. De Natale. Content-based synchronization for multiple photos galleries. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1945–1948. IEEE, 2012.

[5] N. J. Durr, G. González, and V. Parot. 3d imaging techniques for improved colonoscopy. *Expert review of medical devices*, 11(2):105–107, 2014.

[6] D. Hong, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. 3d reconstruction of virtual colon structures from colonoscopy images. *Computerized Medical Imaging and Graphics*, 38(1):22–33, 2014.

[7] G. Kim and E. P. Xing. Jointly aligning and segmenting multiple web photo streams for the inference of collective photo storylines. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 620–627. IEEE, 2013.

[8] J. Kim. Events as property exemplifications. In *Action theory*, pages 159–177. Springer, 1976.

[9] M. Lux and M. Riegler. Annotation of endoscopic videos on mobile devices: a bottom-up approach. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 141–145. ACM, 2013.

[10] B. Munzer, K. Schoeffmann, and L. Boszormenyi. Detection of circular content area in endoscopic videos. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, pages 534–536. IEEE, 2013.

[11] B. Munzer, K. Schoeffmann, and L. Boszormenyi. Relevance segmentation of laparoscopic videos. In *Multimedia (ISM), 2013 IEEE International Symposium on*, pages 84–91, Dec 2013.

[12] G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 23. ACM, 2012.

[13] T. Reuter and P. Cimiano. Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 22. ACM, 2012.

[14] M. Riegler, M. Lux, V. Charvillat, A. Carlier, R. Vliegendhart, and M. Larson. Videojot: A multifunctional video annotation tool. In *Proceedings of International Conference on Multimedia Retrieval*, page 534. ACM, 2014.

[15] M. Zeppelzauer, M. Zaharieva, and M. Del Fabro. Unsupervised clustering of social events. In

# Paper X

# 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 $\rho$ 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. $\rho$ around $0$ or a positive $\rho$ 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

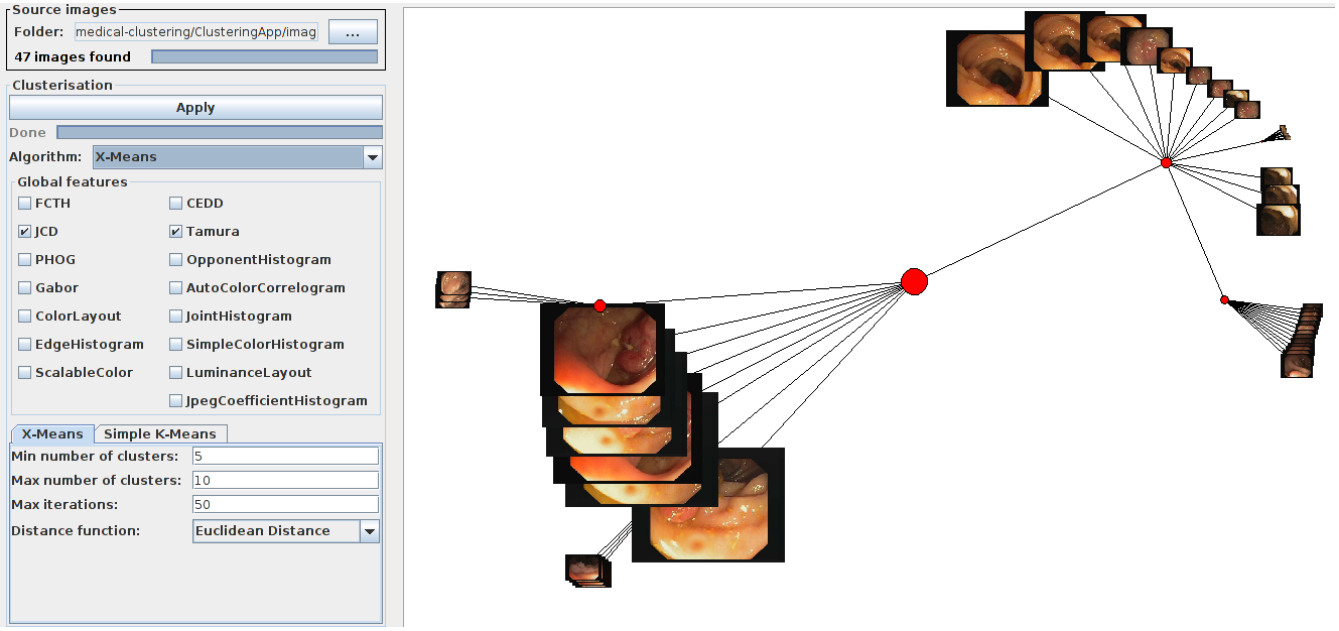



Fig. 1. Demo system: The *left* part contains the settings for the users, and the *right* part shows the output of the clustering as a hyperbolic tree.

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 $\rho$. 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 $\rho$ 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.

[25] M. Lux, "Lire: open source image retrieval in java," in *Proc. of ACM MM*, 2013, pp. 843–846.

# Paper XI

# 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.

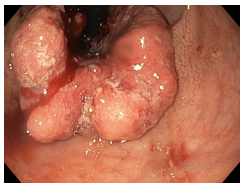


Fig. 2. Colorectal cancer that can be found using colonoscopy.

## 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.

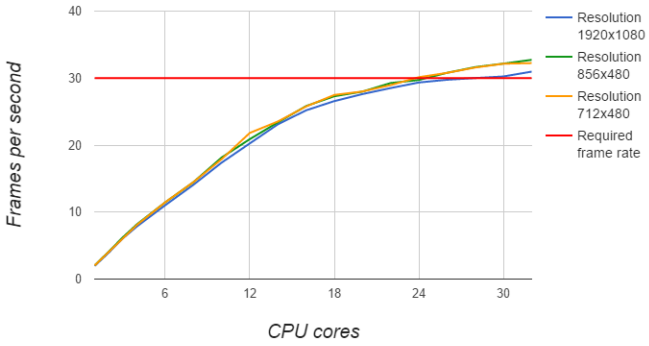



Fig. 3. Processing performance in frames per second.

## 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 XII

# 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, Simula Research Laboratory and University of Oslo
Konstantin Pogorelov, Simula Research Laboratory and University of Oslo
Sigrun Losada Eskeland, Department of Medical Research, Bærum Hospital, Vestre Viken Hospital Trust
Peter Thelin Schmidt, Karolinska Institutet, Dept. of Medicine, Solna and Karolinska University Hospital
Zeno Albisser, Simula Research Laboratory and University of Oslo
Dag Johansen, The Arctic University of Norway
Carsten Griwodz, Simula Research Laboratory and University of Oslo
Pål Halvorsen, Simula Research Laboratory and University of Oslo
Thomas de Lange, Vestre Viken Hospital Trust, Bærum Hospital and Cancer Registry of Norway

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.

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

Additional Key Words and Phrases: Interactive, Medical Multimedia System, Gastrointestinal Tract, Performance, Evaluation

**ACM Reference Format:**
Michael Riegler, Konstantin Pogorelov, Sigrun Losada Eskeland, Peter Thelin Schmidt, Zeno Albisser, Dag Johansen, Carsten Griwodz, Pål Halvorsen, Thomas de Lange, 2015. From Annoation to Computer Aided Diagnosis: Detailed Evaluation of a Medical Multimedia System. *ACM Trans. Embedd. Comput. Syst.* 9, 4, Article 3999 (March 2010), 23 pages.
DOI: 0000001.0000001

## 1. INTRODUCTION

The gap between human users and computer devices such as sensors or cameras has become 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 towards an interdisciplinary research area that combines the medical and multimedia research fields. 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 must target an analysis system for endoscopies that can be used both as a live computer aided diagnostic system and as a scalable detection system for a novel in line screening system using wireless video capsule endoscopes (WVCs).

This work is founded by the Norwegian FRINATEK project "EONS" (#231687).
Contact Author's address: Michael Riegler, Simula Research Laboratory, Norway, email: michael@simula.no;

DOI: 0000001.0000001

The GI tract (shown in figure 1) can potentially be affected by a wide range of diseases. For example, 3 of the 6 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% [World Health Organization - International Agency for Research on Cancer 2012]. These diseases, as well as findings associated with normal (benign) or man-made (iatrogenic) lesions are frequently visualized with endoscopes. The most common dangerous (malignant) lesions are gastric- and colorectal cancer, which are lethal diseases 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 recommend population cohort (everyone above 50 years) is the most important tool for early detection. However, current methods have limitations regarding sensitivity, specificity, access to qualified medical staff and overall cost.

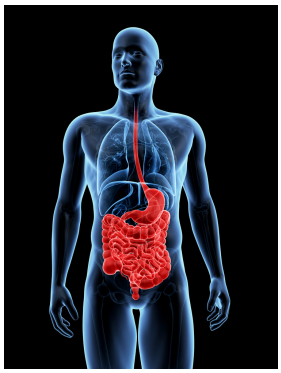
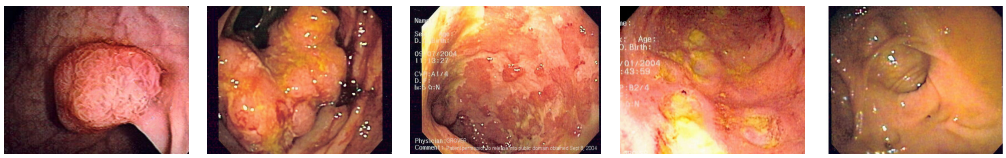


Fig. 1. The GI tract (Image: Kaulitzki/shutterstock.com).

To aid and scale endoscopic examinations, we have developed EIR, named after a Goddess with medical skill in Scandinavian mythology. Eir is 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. 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 WVCs. 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. This paper is an extended version of the paper [Riegler et al. 2016] submitted to the IEEE CBMS conference where we gave a first overview of the system. We extended the paper with a more detailed description of the annotation and visualization subsystem, evaluation and a performance comparison with other system at a grand challenge for endoscopic video analysis.

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 focus for most existing research. Our experiments are therefore limited to one limited scenario. We chose colorectal polyps, a potential precursor for colorectal cancer (CRC), because as statistics show, the life time risk of getting CRC, the second most common cancer for both genders, is 6% [Ferlay et al. 2013]. Obviously, both high precision and recall are of crucial importance, but so is the often ignored system performance in order to provide live feedback. The most recent and most complete related work is the polyp detection system Polyp-Alert [Wang et al. 2015] 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, like for example 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 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 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 decent amount of resources regarding memory consumption and CPU usage which makes it perfectly scalable with more data and different disease to detect in parallel at run time.



(a) Colon Polyp (b) Colorectal Cancer (c) Ulcerative Colitis (d) Crohn's Disease (e) Diverticulosis

Fig. 2. An inconclusive list of abnormalities that can be found using colonoscopy (images from Wikimedia Commons).

The rest of the paper 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

Figure 1 gives an overview of the complete GI tract. This complex system can be affected by various diseases where CRC is one of the most important and a major health issues world wide. Some examples of these diseases and their complexity can be seen in figure 2. If CRC is detected at an early stage, the prognosis is substantial improved, from a 90 percent survival probability for the next 5 years in the early stage 1 to only 5-10 percent 5 year survival probability in the latest stage 4 [Brenner et al. 2016]. Several studies have shown that large population-based screening improves the prognosis and even reduce incidences of CRC [Holme et al. 2013]. Therefore, the current European Union guidelines recommend screening for CRC for all persons older than 50 years [von Karsa et al. 2012].

GI endoscopies are common medical examinations where the lumen and the mucosa of the entire GI tract are visualized to diagnose diseases [Mallery and Dam 2000]. 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 with a 300 W Xenon light bulb. The video signals are transferred to a High Definition LCD screen. The most common, gold standard GI endoscopic examinations are gastroscopy and colonoscopy. However, such endoscopies are demanding 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 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 US dollars (USD$1,100$/person) [The New York Times 2013], and with a time consumption of about one medical-doctor-hour and two nurse-hours, per examination. Furthermore, colonoscopy is not the ideal screening test, and in average, 20% of polyps are missed or incompletely removed meaning that the risk of getting CRC largely depend on the endoscopists ability to detect polyps [Kaminski et al. 2010]. We therefore aim for a system that detects mucosal pathologies in videos of the GI tract where the idea is to assist endoscopists during live examinations.

Moreover, alternatives to traditional endoscopic 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, at least if the examinations should be scaled to everyone above 50, is that endoscopists 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, i.e., enabling larger-scale screenings, and limiting and reducing the traditional endoscopy examinations to patients with positive findings from the WVC examination.

Thus, with EIR, we research and develop a video analysis system that can be used both as a live computer aided diagnostic system and as an automatic detection system for screening systems using

WVCs. As a first step, we target detection of colorectal polyps (see for example figure 2(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 prevent 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 SYSTEM BASIC IDEA

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. Figure 3 gives an overview of the whole system. It can be split into three main parts: the annotation subsystem, the detection and automatic analysis subsystem and the visualization and computer aided diagnosis subsystem. All three parts are important to achieve a system that can support doctors in disease detection and diagnosis in the GI tract.

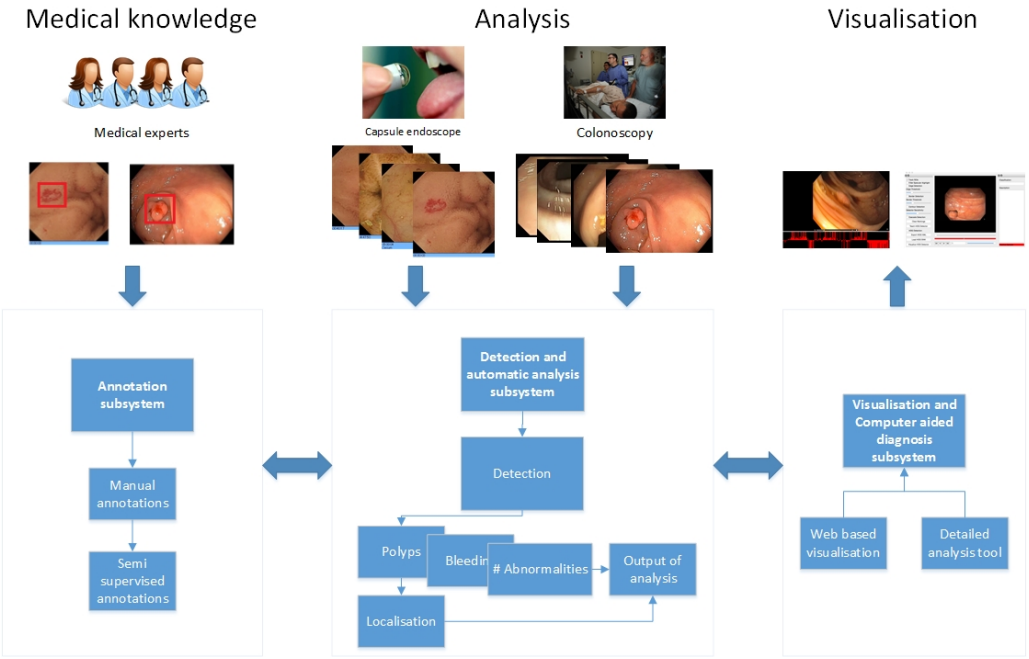



Fig. 3. This is an overview of the whole system. It shows the three main subsystems, namely annotation, detection and automatic analysis and visualization and how they work together. The annotation subsystem extracts and provides the medical knowledge for the detection and analysis subsystem which provides the output to the visualization subsystem.

### 3.1. Annotation Subsystem

The annotation subsystem is used to collect training data for the detection and automatic analysis subsystem. The collection is done with the help of medical experts. It is well known that training data are very important parts of a good classification system. Nevertheless, in the medical field, the time of the experts and multimedia data are two resources that are quite limited. This is primarily because of high every-day workload for physicians, but also due to legal issues. In many countries, patient consent has to be collected before research can be done on each images or videos, 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 [Giritharan et al. 2008]. To aid the system learning process, our new annotation subsystem is an improved version of our semi-automatic annotation tool [Albisser et al. 2015]. The annotation subsystem is the entry point into our whole system, and without this part, the rest of the system is not efficient. The annotation subsystem will makes it easy for doctors to annotate and provide data to the system. The manual annotations are combined with automatic methods that create the final training data.

For example, in a WVC procedure, there are about $216,000$ images per examination, and an endoscopist frequently needs $60$ minutes and even up to two hours to view and analyze all the video data [Li and Meng 2009]. Due to this limitation, it is important to develop automatic methods able to reduce the burden on physicians' and speed up the screening process.

To reduce the amount of time physicians need to spend in the whole process they only have to provide annotations in a single frame of the video or image series. The specialist's knowledge is 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. Because the medical doctor is usually located in a hospital with internal system limitations in terms of what can be installed on the computers by whom and a general usage of old hardware, the implementation of the software is done with standard web technologies which do not require any installation at the hospitals systems.This also includes storage of all information on the systems side and moving the responsibility of maintaining the system and data integrity from the user to the system. Besides getting data for the EIR system to enable automatic screening, the annotation subsystem also makes it possible to use the annotated videos in a medical video archive for procedure documentation or teaching purposes.

The annotation subsystem is divided into manual annotation and object tracking. This is mainly to reduce the time specialists spend on the process because they only have to provide annotations in a single frame. We do require the specialist's knowledge during the first step to do a very basic identification of irregularities and to tag them accordingly. The manual annotation part is to precisely select any regions of interest in a video sequence. The object tracking part is to track the regions of interest on previous and subsequent frames, based on the previously manually created tags. This step is more about tracking an object and adjusting the size and position of the tracked region than about identifying or recognizing irregularities. A specialist's knowledge is therefore not required for the second step. Another reason to divide the process into these two steps is the technologies available for implementing the required software. A specialist is usually located in a hospital with access restrictions due to sensitive patient information. As mentioned before deployment of software is usually difficult in such environments. Nevertheless, internet access and a browser are usually available. This makes standard web technologies a convenient way of circumventing deployment related issues for the manual annotation software. It also implies storing all information on the server side and moves the responsibility of maintaining the system and data integrity from the user to the server administrator.

The manual annotation is the first step in the data gathering process. In this step, a specialist uses rubber band selection (marks a bounded area) to create a coarse selection of regions of interest and annotates every selection with a name for classification. Every region needs to be marked once only. To keep the specialist's time spent on this task minimal, we do not require the region to be marked in the very first video frame it appears. Information on first appearance and change of shape or position within the picture will be added later using object tracking and manual correction. This approach allows a rather rapid way of working for the specialists. They might even watch the video at a higher playback speed and only stop or slow down the playback when necessary. The information collected in this step includes the position and dimensions of irregularities in pixel coordinates, a classification and a timestamp relative to the beginning of the video for each selected region. The annotation component is implemented using JavaScript and HTML5 video, which is available in most recent web browsers.

The output from the manual annotation contains a single tag for every region of interest in the video sequence or images. Using this information, we can now apply object tracking algorithms and manual correction to generate a complete data set. Most of the work in this step is done by the software. The user just needs to step 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. The video must also be rewinded to track the region towards the beginning of the video, as the irregularity most likely has not been marked on the very first frame where it appears. There is of course still a fair amount of manual work involved in this task. However, using a suitable tracking algorithm substantially reduce the time needed to create a complete dataset. Moreover, specialist skills are usually no longer required here as the whole task is simply about tracking regions and adjusting rectangular dimensions rather than actually detecting or recognizing irregularities. The output generated in this step is a list of frames for a certain disease including rectangles for every previously marked region within the frame. Every rectangle in such a list is described by the index of the video frame it belongs to, its position in pixel coordinates and its dimensions. The annotated frames are pooled together as positive and negative samples, which can be used directly in the detection and automatic analysis subsystem.

### 3.2. Detection and Automatic Analysis Subsystem

**Detection Subsystem.** The detection subsystem analyzes videos and images to see if there is anything abnormal to be found in the colon. All the frames that we process in this subsystem 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 or submodels. 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, and because we are not interested in the exact position at this point of the system. In previous work, we showed that global features indeed can outperform or at least reach the same results as local features [Riegler et al. 2014].

The whole system is built using the Lire [Lux 2013] open source library for content-based image retrieval, written in *Java*. 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 WVC videos. Lire uses *Lucene*[1] indices for storing and searching image feature data [Foundation 2013]. Lucene indices are structured in documents, fields and terms. An index contains a sequence of documents, where a document is a sequence of fields, a field is a sequence of terms and a term is a string [Foundation 2013].

The basic idea of our detection subsystem is based on an improved version of a search based method for image classification presented in [Riegler et al. 2014]. 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 like for example deep learning. 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 realized with two separate tools, an indexer and a classifier. We have released the indexer and the classifier as a separate project called *OpenSea*[2], under the terms of the GPL version 3[3].

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

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 indices are then used as input data for the search based classifier. The indexer is created as a separate tool and in a way that it is easy to distribute over different nodes via for example Apache Storm. 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. It creates indices for all directories passed on from the system. The visual features to calculate and store in the indices can be chosen based on the abnormality, because different types of disease require different set 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 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 subsystems 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. Based on this, a decision is made. We refer to these previously generated indices, which are searched for similar image features, as classifier indices or indices 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 an HTML page with a visual representation (see figure 4) of the results, once the processing is finished and a benchmarking function that will output evaluation information (bottom part of figure 5(a)). For the classifier to provide correct benchmarking data, the input data indices must contain either negative or positive samples only, or must have the sample type encoded in the file names of the indexed images. The classifier is parallelized, and one may choose how many CPU cores to use to process the data. A GPU implementation is also currently being developed, and initial experiments show that performance may be imporved further.

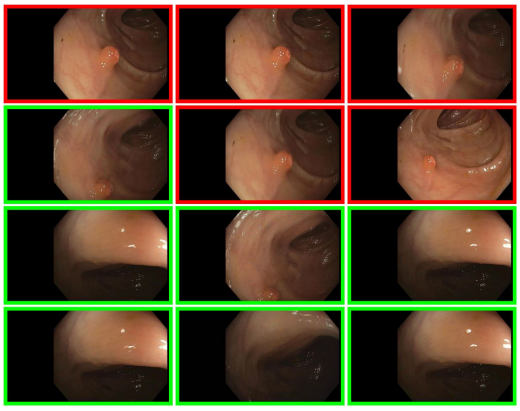


Fig. 4. Results output of the detection subsystem using the features JCD and Tamura. One can see that the detection system could almost always find the polyp containing frames. The first image on the second row is an example for a false negative result.

**Localisation Subsystem.** The detection subsystem cannot determine the location of the detected irregularity. This is done with help of the localization subsystem where the output of the detection system is used as input to determine the exact position of the disease or abnormality. At the moment, it supports the localization of polyps, but it is built in a way that it can be extended to any other automatic detectable disease in our system. The exact positions will be used for the live system,

```
using 4 threads for classifying.
...
image_hortVD_wp_68_71.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_8.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_79.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_82.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_78.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_77.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_81.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:NEGATIVE
image_hortVD_wp_68_83.jpg -> Tamura:NEGATIVE LateFusion:NEGATIVE JCD:POSITIVE
image_hortVD_wp_68_80.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_9.jpg -> Tamura:POSITIVE LateFusion:POSITIVE JCD:POSITIVE
image_hortVD_wp_68_102.jpg -> Tamura:POSITIVE LateFusion:NEGATIVE JCD:NEGATIVE
image_hortVD_wp_68_84.jpg -> Tamura:NEGATIVE LateFusion:NEGATIVE JCD:POSITIVE
image_hortVD_wp_68_101.jpg -> Tamura:POSITIVE LateFusion:NEGATIVE JCD:NEGATIVE
image_hortVD_wp_68_103.jpg -> Tamura:POSITIVE LateFusion:NEGATIVE JCD:NEGATIVE
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Feature     TP   TN  FP  FN  Precision Recall    TNRate    FPRate    Accuracy  FMeasure WFMeasureMccMeasure
Tamura      165  34  37  21  0.816832  0.887097  0.478873  0.521127  0.774319  0.850515  0.850515  0.399001
LateFusion  172  51  20  14  0.895833  0.924731  0.718310  0.281690  0.867704  0.910053  0.910053  0.661481
JCD         162  45  26  24  0.861702  0.870968  0.633803  0.366197  0.805447  0.866310  0.866310  0.509303

writing html output to: results-1435447137.html
duration: 78.293seconds.
```

(a) Console output of the detection subsystem using the features JCD and Tamura.

```
user@user-media-dep: ~/EIR

user@user-media-dep:~/EIR$ ./localizer.sh
CPUs: 4
Using 4 threads
Display 0: 2048x1152
Processing images...
  src_id ;obj_n_lim;   TP   ;   TN   ;   FP   ;   FN   ;precision;  recall ; F-score ;rejected
      all;        1;   1433;      0;    3035;    3069;   0.3207;   0.3183;   0.3195;   15046
      all;        2;   1981;      0;    6791;    2521;   0.2258;   0.4400;   0.2985;   15046
      all;        3;   2333;      0;   10754;    2169;   0.1783;   0.5182;   0.2653;   15046
      all;        4;   2617;      0;   14798;    1885;   0.1503;   0.5813;   0.2388;   15046
      all;        5;   2784;      0;   18975;    1718;   0.1279;   0.6184;   0.2120;   15046
      all;        6;   2900;      0;   23215;    1602;   0.1110;   0.6442;   0.1894;   15046
      all;        7;   2995;      0;   27471;    1507;   0.0983;   0.6653;   0.1713;   15046
      all;        8;   3075;      0;   31774;    1427;   0.0882;   0.6830;   0.1563;   15046
---
 src_size ;frame_time ms; n_frames ; time_total s
       all;      945.302;     19514;    18446.627
   384x288;      824.840;       612;      504.802
   712x480;        0.000;     13500;        0.000
   856x480;     2633.184;      2435;     6411.804
   960x540;     3886.087;      2967;    11530.021
Done
Processing time 4754.769 sec
user@user-media-dep:~/EIR$
```

(b) The final output of the localisation subsystem for polyps. The localiser can easily be extended to localise different disease detect by the detection subsystem.

Fig. 5. System output for the detection and localisation subsystem after the analysis. It includes general results per frame and all evaluation metrics that are provided by the system.

archiving, and size determination. The processing of the images is implemented as a sequence of intra-frame pre- and main-filters, and can easily be extended to localize different diseases (see figure 5(b)) for a console output example). 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 can also be partially hidden and covered by biological substances, like for example seeds or stool, and lighted by direct or ambient light. Moreover, the image itself can be interleaved, noisy, blurry and over-/under-exposed, and it can contain borders and sub-images. Additionally, it can have various resolutions depending on the type of endoscopy equipment or WVC 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 negatively affect the local features 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. These analyses are RGB to YCbCr color space conversion, removal of borders and sub-images, flares masking and low-pass filtering. After the pre-filtering, the images are used for the following local features analysis.

As described above, we have implemented the detection of colon polyps using our local features approach. The main idea of this 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 underlying surface with legs varying in their thickness. These polyps can be approximated with an elliptical shape

region that consists 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 filtered binary image approximated by a set of ellipses from which we build energy maps based on the ellipsis size and border points precision approximation and matching. The final coordinates of one or more polyps in the frame are chosen by looking for maximums in the energy map. An example for the output is shown in figure 6.

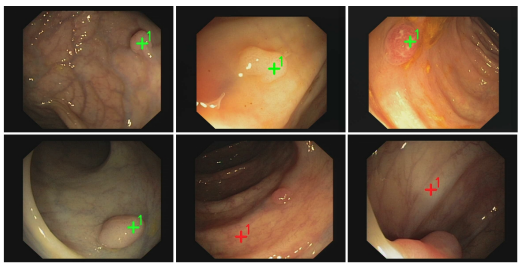


Fig. 6. Output of the localisation subsystem marking the possible locations of polyps. The first 4 frames show an exact match the last two show false positives.

### 3.3. Visualization and Computer Aided Diagnosis Subsystem

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 also under stressful situations that can occur during a live examination. Furthermore, it has to support easy searches and browsing through a large amount of data. This is especially important for the examinations via WVCs.

In general, this subsystem 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 subsystem [Albisser et al. 2015] 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 systems detection and localization part and creates a web based visualization which is then combined with a video sharing platform based on our finding in [Stensland et al. 2014], where doctors are able to watch, archive, annotate and share information. The information collected can later also be used for reinforcement learning for the detection and automatic analysis subsystem.

## 4. SYSTEM EVALUATION

We tested the whole system in terms of detection accuracy and system performance. We also participated at a polyp detection challenge with our system. For any of the subsequent measurements, we used the same computer which was an *old* 32 AMD CPU cores Linux server with 128GB ram. As we will show in section 4.4.3, newer hardware leads to better performance, but the evaluation shows that even with old hardware the system performs as intended. We used the ASU-Mayo Clinic polyp database as training and test data[4]. This dataset is the biggest publicly available 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 [Tajbakhsh et al. 2015].

[4]http://www.polyp2015.com/wp/

### 4.1. Detection Accuracy

For all detection and localization accuracy experiments, we used the common standard metrics precision, recall and F1 score. 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, 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 [Efron and Tibshirani 1997].

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 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. 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. For our first test, 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 test are presented in table I. The single image feature that generally achieves the best score is CEDD [Lux 2013]. All features used here are described in detail in [Lux and Marques 2013]. Further, the image features JCD, EdgeHistogram, Rotation Invariant Local Binary Patterns, Tamura and Joint Histogram achieve promising results. 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 as presented in table II, we have decided the image features JCD and Tamura seem to be the best ones for our performance measurements because they have a good precision and recall but at the same time the computation time is low.

To assess the actual performance of the classifier using these two image features, we conducted a leave-one-out cross-validation with all available video sequences, and 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. average weighted f-score value of $0.929$.

### 4.2. Localization Accuracy

For this experiment, we used the ground truth provided in the ASU-Mayo clinic polyp dataset. This ground truth contains the exact positions of the polyps in the frame. Table IV present the performance of the localization subsystem evaluation. The subsystem 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. One problem we are currently working on is that the localization outputs four possible disease po-

Table I. Leave-one-out cross-evaluation combined for all supported features.

| Feature | True Positive | True Negative | False Positive | False Negative | Precision | Recall | F1 score |
|---|---|---|---|---|---|---|---|
| JointHistogram | 3,369 | 13,826 | 1,085 | 511 | 0.7563 | 0.8682 | 0.8084 |
| JpegCoefficientHistogram | 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 |
| FuzzyOpponentHistogram | 3,341 | 13,552 | 1,359 | 539 | 0.7108 | 0.8610 | 0.7787 |
| SimpleColorHistogram | 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 |
| FuzzyColorHistogram | 2,708 | 13,243 | 1,668 | 1,172 | 0.6188 | 0.6979 | 0.6560 |
| RotationInvariantLBP | 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 |
| LocalBinaryPatternsAndOpponent | 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 |
| OpponentHistogram | 2,246 | 14,157 | 754 | 1,634 | 0.7486 | 0.5788 | 0.6529 |
| EdgeHistogram | 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 |

Table II. Top $20$ results of feature combinations. Each combination contains 2 image features for the video wp_61, sorted by F1 score.

| Feature | True Positive | True Negative | False Positive | False Negative | Precision | 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 |

sitions per frame. At least one of them points at the polyp in all cases, but for the evaluation all four points where included in the calculations, which influences the performance metrics negatively.

### 4.3. MICCAI Challenge

To see how our method compares to other state of the art methods, we participated in the Endovis Automatic Polyp Detection in Colonoscopy Grand Challenge [5] at the 2015 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). The challenge was

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

Table III. 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 |
| $1,920 \times 1,080$ | 908 | 1 | 0 | 0 | 1 | 1 | 1 |
| wp_24 | 310 | 68 | 127 | 12 | 0.7093 | 0.9627 | 0.8168 |
| $1,920 \times 1,0809$ | 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 |

Table IV. Performance evaluation of the localization algorithm.

| Data set | 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 |

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 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 for the challenge since our system is not built for polyp detection only. Other participants used a wide range of different methods to detect polyps. These methods ranged from hand crafted features like for example 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 for example 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 the RUS approach

Table V. Results of the MICCAI polyp localisation challenge.

| 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 VI. Results of the MICCAI polyp detection challenge. The table shows the detection latency in milliseconds and F1 score.

| 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 |

used a deep learning method. Based on the fact that our system is not built for only polyp detection, the results are 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, etc. 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 polyp. All in all, the results of the challenge are good for a system that is designed to be extendible and refine able for different disease. 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

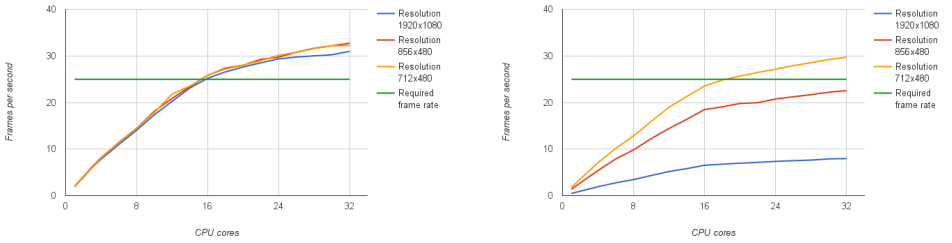
One further requirement for the system is scalability and performance. The idea is to use the system to do mass screening for lesions in the GI tract, using video sequences recorded live with colonoscopy or WVCs. For the evaluation, we decided to use the configuration of the system that performed best in the accuracy experiment. For the detection system, this is a combination of two features (JCD and Tamura). This is the scenario that that will be used in the live system tests later. Therefore, it is important to reach real-time performance in terms of processing a video and reach a frame rate of not less than $25$ frames per second. For all the tests, we used three videos from three different endoscopic devices and different resolutions. The three videos are wp_4 with a resolution of $1,920 \times 1,080$ and $910$ frames, wp_52 with $856 \times 480$ and $1,106$ frames and np_9 with $712 \times 480$ and $1,843$ frames. We chose these videos to show the performance under the different requirements that the system will have to face when it is used.

*4.4.1. CPU Processing.* To test how the different parts of our system scale in terms of used CPU cores, we performed several tests on our test machine. For all tests, we measured time per frame for the number of used cores. We also conducted some experiments to understand the influence of the size of the training data on the performance.

For the detection approach, we first measured the indexing part that creates the model that is later on used by the classifier. This process does not have to be in real-time and can be seen as batch processing but it should at least be scalable for larger datasets. We did all experiments on a Linux machine using $16$ CPU cores. Extracting two features and indexing them for the whole ASU Mayo dataset takes in 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 VII shows the overall results. We found that a larger

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

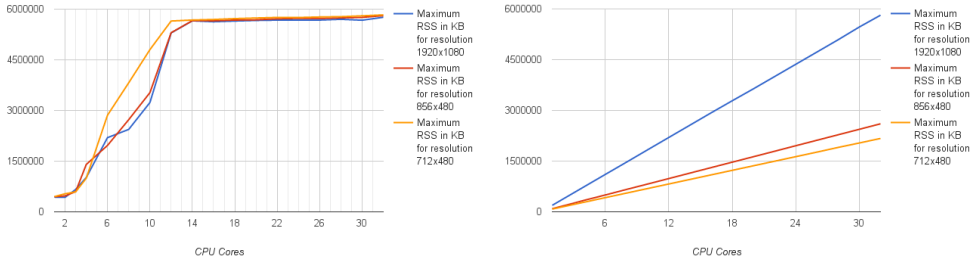
| Index | frames | total time in seconds | time per frame in ms |
|---|---|---|---|
| $D1$ | $3,871$ | $89.78$ | $23.1$ |
| $D2$ | $14,909$ | $178.55$ | $11.9$ |
| $D3$ | $29,818$ | $231.75$ | $7.7$ |
| $D4$ | $100,000$ | $782.351$ | $7.8$ |




(a) The detection subsystem performs efficiently and the required frame rate is reached with 16 CPU cores used in parallel.

(b) The localisation subsystem performance depends heavily on the resolution of the videos.

Fig. 7. Detection and localisation subsystem performance in terms of frames per second depending on the number of CPU cores and the resolution of the videos. The resolutions are $1,920 \times 1,080$, $856 \times 480$ and $712 \times 480$.




(a) This chart shows the overall memory consumption for all three videos in the detection part. A maximum is reached at around 14 videos in the detec- tion part is scalable.

(b) This chart shows the overall memory consumption for all three videos in the localisation part. This shows us that the localisation used CPU cores. Further investigation is needed to see if the part scales well in terms of memory.

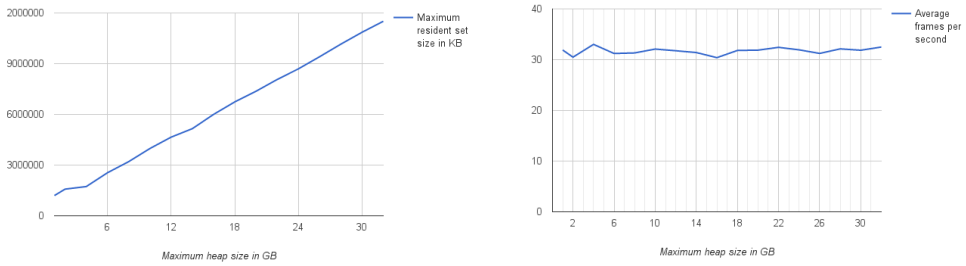
Fig. 8. Overall memory consumption of the detection and localization subsystem for the three test videos.

dataset leads to a faster indexing time per frame. We think this is due to the Java runtime optimizer. Furthermore, we did not find an increase after more than $30,000$ frames in the dataset. We think that 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 be able to meet all requirements of the system for future tasks. The performance of the detection is more important, since the system should provide a result as fast as possible and not slower than 25 frames per second making it usable for live applications. For all tests, we used the 3 different videos described before.

Figure 7(a) shows the detection subsystem performance for the three tested resolutions in fps. The required frames per second for all three videos are reached with 16 CPU cores used in parallel.

Figure 7(b) shows the performance of the localization subsystem for FPS. For the highest resolution, namely $1,920 \times 1,080$, the best result is 7.9 fps. A significant code optimization and using of GPU for accelerated calculations will be needed to reach required fps. For resolution $856 \times 480$ the required frames per second are almost reached with 32 CPU cores used in parallel. The best result is 22.5 fps for this resolution. A slight code optimization will be needed to reach the required fps. For the final video with the resolution $712 \times 480$, the required frames per second are reached with 19 CPU cores used in parallel. The outcome of these experiments for the localization subsystem clearly shows that our system can reach real-time requirements but needs some optimization. For the detection subsystem the required frames per second are reached in all resolutions.

*4.4.2. Memory.* In the memory experiments, we tried to find out how the different parts of the system scale in terms of memory. Moreover, we had a look into the influence of the index sizes on the performance. The memory usage for both subsystems is shown in figure 8(a) and figure 8(b). In the




(a) To understand the memory consumption of the detection part we had a closer look into the Java garbage collector. It showed us that it always used all the memory that it can get. Automatically it is set to around 6GB on our system.

(b) This experiment showed that the available memory for the detection part does not influence the frames per second performance. The Java memory scheduler takes always the whole memory that it can get but it also works perfectly with only 1GB. This is a proof that the detection part is not dependent on memory and therefore memory is not a bottleneck for scaling the system.

Fig. 9. Memory usage experiment for the detection subsystem. Firstly, we tested how the maximum heap size influences the memory usage. Secondly, we investigated which influence the heap size has on the frames per second performance of the detection part.

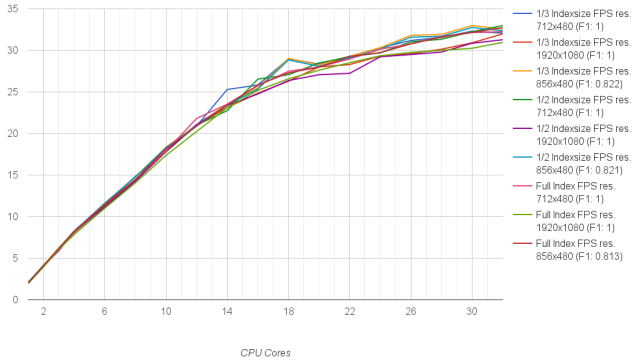



Fig. 10. This chart shows how the amount of training data influences the performance of the detection subsystem in terms of detection accuracy and frames per second output. The training data has been reduced to 1/2 of the original size (ca. $8,800$ frames) and 1/3 (ca. $5,800$ frames). The chart shows that there is no significant difference for the detection performance and the frames per second. The smaller indexes can achieve even a better F1 score for $856 \times 480$.

localization, the memory usage behaves as expected (linear growth) and shows that the localization is scalable in terms of memory. For the detection subsystem the memory usage behaves rather unusual after a certain number of used CPU cores. Therefore, a closer look into it was necessary. The results of this closer look can be found in figure 9(a) and 9(b). We tested different memory sizes used for the detection starting from 1GB up to 32GB. These tests showed, 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 is not dependent on memory, and therefore, memory is not a bottleneck for scaling the system.

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 10 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 frames per second output of the detection system. It is possible that

an index with several hundred thousands frames will most probably lead to a lower frames per second output. But, in the medical field, for that the presented system is intended, 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 $856 \times 480$ the F1 score increased slightly compared to the full training data. This shows that the detection subsystem also performs very well with a smaller amount of training data, which is a very positive point for the medical field because of the constant lack of training data.

*4.4.3. Cloud Experiments.* All the evaluations presented before have been performed on rather old hardware (4 years). To get an idea of how the performance would be on an actual hardware setup for the detection subsystem, we conducted tests on several Amazon AWS EC2 instances. For all the tests, we used one *c4.8xlarge* instance and one *c4.4xlarge* instance.

On the *c4.8xlarge* instance (Intel Xeon E5-2666 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 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

As we aim for a system to be used both as a live computer aided diagnostic system and as a scalable detection system WVC videos, we are currently working on two different real world use cases with our partner hospitals. The first one is a live system that should support and assist endoscopists while they do live examination. The second one has as goal to automatically analyze videos captured by WVCs. The live system requires fast and reliable processing, and the WVC 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 analysis to the doctor. In the future, if the system is well tested under clinical conditions, it can be combined in one screen. Endoscopy is a common GI examination and is essential for the diagnosis of most mucosal diseases in the GI tract, particularly diagnosis of CRC and its precursors. Previous studies have demonstrated that a major challenge is the detection rate of lesions [Tanaka et al. 2013; de Lange et al. 2005]. 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. At the beginning, we plan to optically 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 paper [de Lange et al. 2005] shows, single images can be an efficient way to store important findings from an examination.

### 5.2. Wireless Video Capsule Endoscope

The multi-sensor WVC is swallowed in order to visualize the GI tract for subsequent detection and diagnosis 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.

The current WVCs have a low resolution of $256 \times 256$ with 3-30 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 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 colonoscopies also for WVCs [Khaleghi and Balasingham 2015; Chandra and Balasingham 2015].

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, and distributed processing researchers in order to scale the big 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 leading 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 and to the best of our knowledge such a system does not exist at the moment. In our case it has to be able to process large amount of data efficiently and in real-time, it has to be a complete system that can be used in a real scenario, it should support doctors live during colonoscopies, it should be able to process data from WVC and manual colonoscopy and finally it should be expendable with different disease and input data. All this requirements touch different areas of related work. In the following, we will discuss the most relevant works for our proposed EIR system.

### 6.1. Annotation

To build a ground truth dataset, it is necessary to collect video sequences with frames that show an object to be detected (positive frames), as well as frames where no such object is present (negative frames). It is then necessary to select the region on the positive frames where such an object is visible. Such selection builds a dataset containing both positive and negative samples. To create such a dataset, appropriate tools are required.

Liu et al. [Liu et al. 2007] 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, uploading videos to a storage server, removing redundant video frames, separating or merging video sequences, segmentation based on audio features (speech recognition), post-processing and analysis of colonoscopic procedures. *Arthemis* was designed to facilitate the process of reviewing videos, locating and annotating important content, and exporting annotated content for research, teaching and training purposes. Arthemis supports annotation by ellipse selection and free-hand-drawing. It is written in Java and C, and uses a third-party MPEG encoding/decoding-library for an extra fast playback mode, as Microsoft's native multimedia toolkit DirectShow was considered not to provide the needed performance. A special feature of Arthemis is that it provides the ability

to view automatically detected segments of colonoscopy videos. Annotated figures are stored in a proprietary format based on xml. The software has been designed to be easy to use by physicians, allow fast playback, and be efficient, robust and extensible. According to the article, the software starts to become a useful tool for endoscopic research and education. It is proposed to be used by Medical students, residents and fellows, for learning to recognize the common endoscopic abnormalities and the therapeutic modalities, used by experienced endoscopists. The user-interface of Arthemis is considered to be intuitive and easy to use. The component-based design of the software is listed as a strength, since it allows multiple developers to develop new components at the same time, without worrying about losing the control of the code. We consider this article very significant, as it describes an existing, complete implementation of an integrated solution for collecting, archiving, processing, annotating and visualizing colonoscopy videos. A potential weakness of the implementation is the use of proprietary components. The deployment requires both a server system and installation on the client side, making the entire solution hard to obtain and distribute. Also, the number of supported features seems rather large. Generally, a large number of features can be considered a strength, it usually also makes a tool less intuitive to use.

Riegler et al. [Riegler et al. 2014] present a web based annotation tool called VideoJot, supporting several different kinds of annotations. The tool is not specifically designed for medical usage. The article researches the benefits of different video annotation features in communicating general concepts of a video game based on captured game sessions. Game play recordings were selected as a domain, because it specifically requires the software to cope with fast camera movements. The tool provides simple and easy to use controls for annotating with temporal and spatial information and functionality to enrich the content of the video with added information. Annotations can be done by free hand drawing on top of the video, either during playback or alternatively, when the video is paused. The annotations will be replayed during video playback; if an annotation was added when the video was paused, the playback pauses for replaying the annotation. An additional feature is the possibility to zoom into the video, i.e., annotations created in zoomed-in mode will automatically be replayed in zoomed-in mode.

For simple temporal annotations, VideoJot further provides LikeLines [Vliegendhart et al. 2012] - a bar below the video, displaying a one-dimensional heatmap. The heatmap displays which parts of the video received explicit ”likes”. The tool is written entirely in HTML5 and JavaScript, using the HTML5 video element for playback and HTML canvas for drawing. VideoJot is interesting for the medical field because it provides a straight forward approach for collecting video sequences and creating annotations, which we reuse for collecting colonoscopy videos from hospitals by providing similar functionality to receive basic annotations from endoscopists. As mentioned in the conclusion for the VideoJot article, users, who tested the software, still considered it too complicated and considered training necessary for creating good annotations. It is also stated that the ability to slow down the playback when creating an annotation is important and that users want text based annotation in addition to the hand drawn one. The zooming function was not considered important, unless the video sequence to be annotated is of high resolution and high level of detail.

The use of an annotation tool for endoscopy videos is further researched by Lux and Riegler [Lux 2013]. 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. A tool for an Android tablet computer is presented, which uses the touch screen, motion sensor and speech recognition for user interaction. This tool is required to be easily integrable into existing business processes in medical information systems. Hence, complicated installation and hardware requirements were not acceptable. A low cost off-the-shelf tablet computer, however, is considered a good choice. The following features are integrated into the software: The video can be manually segmented into non-overlapping pieces, selected segments can be annotated using speech, text annotations can be added by using the integrated Google speech-to-text web service, sketch-like drawings can be added on top of the video, and shaking the device is used for annotating important events in the video sequence. All the annotation information is

kept and stored separately from the video file. Audio recording is stored in a compressed form, and the drawings are stored as path information. As a technical challenge, the paper mentions the drawings which have to be drawn on top of the video overlay. The replay is not accurate in terms of frames, but is considered good enough in terms of accuracy and excellent in terms of robustness and performance.

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 subsystem in EIR builds up on technologies and methods from [Riegler et al. 2014] and [Lux 2013] to reach optimal annotation performance.

### 6.2. Automatic Analysis Systems for 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. Automatically analysis of polyps in colonoscopies has been in focus by research for a long time and several studies have been published [Wang et al. 2013; Wang et al. 2010; Wang et al. 2011]. However, not many complete 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 most cases is polyp detection for a specific type of camera. Several algorithms, methods and partial systems have been proposed and have achieved at the first glance 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 amount 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. [Wang et al. 2015] 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. For classifying video endoscopy imaging data, most approaches relay on using a support vector machine (SVM) or binary classifier in some way, or they could be used as a pre-processing step for them. The features, used for the SVM or the binary classifier vary a lot depending on the approach. Some methods use physical dimensions, grayscale intensity values, gradient orientation, RGB color information or texture as input for the classifier. Each of these approaches has advantages and disadvantages. In general, it can be distinguished between two different approaches for the automatic detection of disease in the colon. These approaches are geometrical analysis and machine learning. They could both be used for imaging data that was recorded with a conventional colonoscope or with a camera capsule. Moreover, it is also possible to use these methods with data from a virtual colonoscopy. However, such data are significantly different from camera recorded data and are not discussed in detail here.

Mamonov et al. [Mamonov et al. 2014] 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 algorithm process separate input frames and classifies each frame as either containing a polyp or not. The assumption is that polyps can be generalized as protrusions (something that bumps out) that are mostly round in shape. This assumption was tested on a dataset created from frames of video sequences of five different patients. Based on these tests, the algorithm reached a sensitivity of $81.25\%$ per polyp at a specificity level 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

Table VIII. Performance comparison of polyp detection approaches discussed in this chapter. Not all performance measurements are available for all methods, and different datasets are used. Nevertheless, including every 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. [Wang et al. 2015] | polyp / edge, texture | 97.70% | – | – | 95.70% | 10 | 1.8m frames |
| Wang et al. [Wang et al. 2014] | polyp / shape, color, texture | 81.4% | – | – | – | 0.14 | 1,513 images |
| Mamonov et al. [Mamonov et al. 2014] | polyp / shape | 47% | – | 90% | – | – | 18,738 frames |
| Hwang et al. [Hwang et al. 2007] | polyp / shape | 96% | 83% | – | – | 15 | 8,621 frames |
| Li and Meng [Li and Meng 2012] | tumor / textural pattern | 88.6% | – | 96.2% | 92.4% | – | – |
| Zhou et al. [Zhou et al. 2014] | polyp / intensity | 75% | – | 95.92% | 90.77% | – | – |
| Alexandre et al. [Alexandre et al. 2007] | polyp / color pattern | 93.69% | – | 76.89% | – | – | 35 images |
| Kang et al. [Kang and Doraiswami 2003] | polyp / shape, color | – | – | – | – | 1 | – |
| Cheng et al. [Cheng et al. 2008] | polyp / texture, color | 86.2% | – | – | – | 0.076 | 74 images |
| Ameling et al. [Ameling et al. 2009] | 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 |

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. [Hwang et al. 2007]. 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. Ellipses are then fitted into the segments by constructing a binary edge map for each segmented region and using a least square fitting method. A threshold-function is used for the creation of the edge map. Regions with too little edge information in their respective edge map are discarded. These ellipses are then further evaluated for matching of curve direction, curvature, edge distance and intensity. The curvature of the ellipse is split into six parts. At least two adjacent parts must have a strong edge pattern, otherwise, the ellipse is discarded. Lumen areas are filtered out by applying a threshold on the intensity of the ellipse. 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. This allows to apply a threshold in number of frames for the detection to reduce the number of false positives. 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 [Mamonov et al. 2014], 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 negatives. The correctness of this assumption depends strongly on the frame rate of the camera that is used for recording the video.

The most recent and complete system in the well researched polyp detection field is Polyp-Alert [Wang et al. 2015] 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 taken from 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 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. Another recent approach related to our approach and not limited to polyps is presented by Nawarathna et al. [Nawarathna et al. 2014]. 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.

Other papers that discuss how to improve performance of endoscopic surgeries in general (not colonoscopy) are for example [Munzer et al. 2013c; Munzer et al. 2013a; Munzer et al. 2013b]. 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, making them interesting for our system. Since neural networks are commonly used nowadays they are also discussed in relation the GI tract analysis.

### 6.3. Deep Learning

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 [Deng et al. 2009]. Nevertheless, they have some negative aspects that make them less useful for this use case [Chin and Brown 2000]. 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 *blackbox* approach, which neural networks are known for, seem 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 [Nguyen et al. 2014]. The best way is still to understand the problem and then solve it. Further, 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. 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 interesting to detect [Schmidhuber 2015]. 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 give a much more accurate probabilistic score that can be directly translated to the real world scenario [Specht 1990].

### 6.4. Summary

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

## 7. CONCLUSION

In this paper, a complete multimedia system for annotation, automatic disease detection and visualization in context of the GI tract has been presented. 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 (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, in our case live colonoscopies, for systems recording videos faster than 30 fps. 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 gave a glance on a real world implementation and use case of our system that is currently being built. This includes analysis of WVC videos and live support of colonoscopies. 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 hospitals all over Europe. 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, etc. Finally, we are working on an extension that allows the system to utilize GPUs to make it even faster.

## REFERENCES


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*.

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

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.

Hermann Brenner, Matthias Kloor, and Christian Peter Pox. 2016. Colorectal cancer. *The Lancet* 383, 9927 (Feb. 2016), 1490–1502. DOI: http://dx.doi.org/10.1016/S0140-6736(13)61649-9

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 IEEE EMBC*. 4093–4096.

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.

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.

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.

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 IEEE CVPR*. 248–255.

Bradley Efron and Robert Tibshirani. 1997. Improvements on Cross-Validation: The .632+ Bootstrap Method. *J. Amer. Statist. Assoc.* 92, 438 (1997), pp. 548–560. http://www.jstor.org/stable/2965703

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), 1374–403.

The Apache Software Foundation. 2013. Apache Lucene - Index File Formats. (2013). https://lucene.apache.org/ Accessed: 2015-07-29.

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*.

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. *Cochrane Database of Systematic Reviews* 9 (2013). DOI: http://dx.doi.org/10.1002/14651858.CD009259.pub2

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.

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.

J Kang and R Doraiswami. 2003. Real-time image processing system for endoscopic applications. In *Proc. of IEEE CCECE*. 1469–1472.

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

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.

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.

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.

Mathias Lux. 2013. LIRE: Open Source Image Retrieval in Java. In *Proc. of ACM MM*.

Mathias. Lux and Oge Marques. 2013. *Visual Information Retrieval Using Java and LIRE*. Vol. 25. Morgan & Claypool.

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

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.

Bernd Munzer, Klaus Schoeffmann, and Laszlo Boszormenyi. 2013a. Detection of circular content area in endoscopic videos. In *Proc. of CBMS*. 534–536.

Bernd Munzer, Klaus Schoeffmann, and Laszlo Boszormenyi. 2013b. Improving encoding efficiency of endoscopic videos by using circle detection based border overlays. In *Proc of ICME workshops*. 1–4.

Bernd Munzer, Klaus Schoeffmann, and Laszlo Boszormenyi. 2013c. Relevance Segmentation of Laparoscopic Videos. In *Proc. of ISM*. 84–91.

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).

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).

Michael Riegler, Mathias Lux, Vincent Charvillat, Axel Carlier, Raynor Vliegendhart, and Martha Larson. 2014. VideoJot: A Multifunctional Video Annotation Tool. In *Proce. of ACM ICMR*. 534:534–534:537.

Michael Riegler, Konstantin Pogorelov, Zeno Albisser, Pål Halvorsen, Thomas de Lange, Dag Johansen, Peter Thelin Schmitdt, Sigrun Losada, and Carsten Griwodz. 2016. EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal Endoscopies. *Submitted to IEEE CBMS 2016* (2016).

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

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

Håkon Kvale Stensland, Vamsidhar Reddy Gaddam, Marius Tennøe, Espen Helgedagsrud, Mikkel Næss, Henrik Kjus Alstad, Asgeir Mortensen, Ragnar Langseth, Sigurd Ljødal, Østein Landsverk, and others. 2014. Bagadus: An integrated real-time system for soccer analytics. *ACM TOMM* 10, 1s (2014).

Nima Tajbakhsh, Suryakanth Gurudu, and Jianming Liang. 2015. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. (2015).

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.

The New York Times. 2013. The $2.7 Trillion Medical Bill. (June 2013). http://goo.gl/CuFyFJ [last visited, Nov. 29, 2015].

Raynor Vliegendhart, Martha Larson, and Alan Hanjalic. 2012. LikeLines: collecting timecode-level feedback for web videos through user interactions. In *Proc. of ACM MM*. 1271–1272.

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.

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.

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.

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.

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).

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 BME* 57, 3 (2010), 685–695.

World Health Organization - International Agency for Research on Cancer. 2012. Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. http://globocan.iarc.fr/Pages/fact_sheets_population.aspx. (2012).

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

# Computer Aided Disease Detection System for Gastrointestinal Examinations

# Computer Aided Disease Detection System for Gastrointestinal Examinations

Michael Riegler[1], Konstantin Pogorelov[1], Jonas Markussen[1], Mathias Lux[3], Håkon Kvale Stensland[1]
Thomas de Lange[2,4], Carsten Griwodz[1], Pål Halvorsen[1], Dag Johansen[5], Peter T Schmidt[6], Sigrun L. Eskeland[4]
[1]Simula Research Laboratory & University of Oslo [2]Cancer Registry of Norway [3]Klagenfurt University
[4]Vestre Viken Hospital Trust [5]Univeristy of Tromsø - The Arctic University of Norway
[6]Department of Medicine Solna, Karolinska Institutet, Center for Digestive Diseases, Karolinska University Hospital
michael@simula.no, mlux@itec.aau.at, t.d.lange@medisin.uio.no, sigesk@vestreviken.no, peter.thelin-schmidt@karolinska.se

## ABSTRACT

In this paper, we present the computer-aided diagnosis part of the EIR system [9], which can support medical experts in the task of detecting diseases and anatomical landmarks in the gastrointestinal (GI) system. This includes automatic detection of important findings in colonoscopy videos and marking them for the doctors. EIR is designed in a modular way so that it can easily be extended for other diseases. For this demonstration, we will focus on polyp detection, as our system is trained with the ASU-Mayo Clinic polyp database [5].

## CCS Concepts

**•Information systems → Multimedia and multimodal retrieval;**

## Keywords

Medical Multimedia; Information Systems; Classification

## 1. INTRODUCTION

Colonoscopy is an invasive medical procedure, where medical experts (endoscopists) investigate and operate on the colon, i.e., by using a flexible endoscopes as shown in figure 1(a). From the tip of the endoscope, a video is transmitted, and the endoscopists rely on the video stream to diagnose disease and apply treatments. As the camera is the virtual eye of the endoscopist and the video stream is all the endoscopist perceives, research in medical imaging focuses on diagnosis and detection of diseases and anatomical landmarks based on video. A video capsule endoscope (VCE) (camera pill see figure 1(b)) is an alternative non-invasive technique to record videos from the colon. The capsule with a camera is swallowed, it records a video of the gastrointestinal (GI) tract, and an endoscopist analyses the videos afterwards for endoscopic findings.



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



ACM ISBN 978-1-4503-4297-1/16/05.
DOI: http://dx.doi.org/10.1145/2910017.2910629

(a) An endoscope as it is used for standard colonoscopies. It consists of a control device, the tube and the camera at the tip.

(b) Examples for two video capsular endoscopes. The upper one has two cameras and the lower has one.

**Figure 1: Two devices that produce endoscopic videos.**

In this paper, we present a demo that shows how the visualization and computer-aided diagnosis part of our system works and how it can be used by medical experts to support them at the time of the colonoscopy procedure, as well as after the procedure has finished. There are several potential benefits of such a system for patients and the healthcare sector. It could be a useful tool for training of new endoscopists in recognizing and classifying endoscopic findings, and probably also improve endoscopists' live detection of polyps. Early detection prevents the polyps from developing into colorectal cancers (CRC), the second most common cancer for both genders with a 6% lifetime risk of contracting the disease. The automatic detection could also be applied to VCE videos, thus eliminating or reducing the time to review the video footage and freeing time for the endoscopist to perform other important medical tasks. The automatic computer interpretation makes it also possible to generate automatic text reports from the procedures, and the patients can receive the results from the examination faster. The detection subsystem is used in combination with our previously developed TagAndTrack tool [1] to be able to provide computer-aided diagnosis to endoscopists. The detection subsystem is released as open source software [1]. The automatic detection of irregularities and the segmentation of videos can help doctors to save time, and can further increase the accuracy of diagnosis and be used to verify the completeness of an examination. Moreover, our system is very easy to train, to modify and to expand so that it can be used and improved by everyone, even by non experts. We also want to point out that our system is not limited to the medical use case. It could be expanded to many different use cases that can be solved by

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

visual content-based classification, like for example content-based segmentation of sport events like soccer games. The remainder of the paper presents the architecture and the implementation of the application. Furthermore, we show how the system can be used. A demo video of the tool can be found at https://youtu.be/gb2BqMuZ2h0. In this demo video, one can easily see the challenges that come with colonoscopy videos, i.e., blurry frames, low resolution, reflections and fluids, etc.

## 2. RELATED WORK

Automatic detection of polyps in colonoscopy has been in focus of research for a long time [12, 14, 11]. 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 [13], which is able to give near real-time feedback during colonoscopies. The system can process 10 frames per second, and visual features and a rule-based classifier are used to detect the edges of polyps. The researchers report a performance of 97.7% correctly detected polyps in their data set. Compared to our EIR system [9], this system seems to reach higher detection accuracy, but our system is faster and can at the moment detect polyps in real-time. Furthermore, our system is not restricted to detecting polyps and will be extensible to detect several different diseases at the same time. To achieve real-time for a multi-class detection, we plan to utilize heterogeneous architectures such as GPUs. Another recent approach that is related to our system is presented by Nawarathna et al. [6]. The authors describe a method to detect abnormalities like bleeding, but also polyps in colonoscopy videos using a texton histogram of an image block. In a nutshell, our system uses global image features for the classification of frames and a search-based approach that leads to low classification times per frame. It is well known that global image features are very easy to extract and analyze in terms of time and easy to store in terms of space. This makes our EIR well suited for applications on huge amounts of data [4].

## 3. ARCHITECTURE & IMPLEMENTATION

Our detection subsystem consists of two modules, (i) an *indexer* and (ii) a *classifier* as shown in figure 2. The *indexer* analyzes input data and extracts global features from the training videos. The *classifier* is in principle a binary K-Nearest-Neighbor (K-NN) classifier, which utilizes the index to search the training set for visually similar cases. The results of multiple global features are fused and weighted by the classifier module and result in a proposed class. The *classifier* works on single frames, but also accepts a complete video as input. In this case, it will classify every single video frame, and it will output a result file. We modified the previously developed *TagAndTrack* tool [1], which can open and interpret the results of the classifier for visualizing the classification results.

We have implemented the *indexer* as well as the *classifier* in Java. We are using *LIRE* [4] for extracting global image features, and *LIRE* internally uses *Lucene*[2] for creating

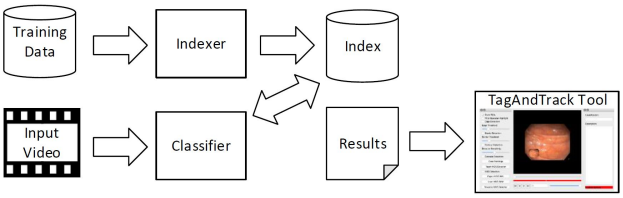



**Figure 2: The overall architecture of our demo. The detection subsystem provides the output that we use in a visualization tool that presents it to the medical experts.**

and searching indexes. Further, we are using *OpenCV*[3] for reading and decoding video files. The *indexer* as well as the *classifier* both use multiple threads.

### 3.1 Detection Subsystem

The classification of each frame is based on the analysis of search results for a given query image. As mentioned before, the classification algorithm is a modified K-NN algorithm. 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 a simple majority decision based on the outcome of the K-NN algorithm. 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 a 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 behaviour of 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 indices that are generated off-line or on-the-fly, based on single or multiple image features. For each image in the input index or video, it searches the provided classifier indices and finds the images with the most similar image features, whereas similarity is determined based on 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 *recall, precision, weighted F1-score, etc.*. For this to work, the input data must be labelled correctly before it is classified. This can either be done by prefixing the file names of the files in the index with 'p' or 'n' for positive and negative samples respectively, or by supplying separate indices with the command line options '-P' and '-N' for the input data.

[2]https://lucene.apache.org/

[3]http://opencv.org/

## 3.2 Detection Subsystem Usage Examples

In the following, some examples are presented of how to use the detection subsystem to classify input videos.

```
java -jar classifier.jar \
-p /pos/index -n /neg/index \
-i /my/index -f JCD -f FCTH
```

This example shows how to classify images from the index */my/index* using the image features *JCD* and *FCTH*, by finding the most similar images among the positive samples from */pos/index* and the negative samples from */neg/index*. For the calculation of the evaluation metrics, it is required that the images indexed in */my/index* have names starting with 'p' or 'n' for positive or negative samples, respectively. This generates visual classification output in HTML format.

```
java -jar classifier.jar \
-p /pos/index -n /neg/index \
-v /my_video.avi -f JCD
```

In our last example, a video file is supplied as input to the classifier. All video frames of this input video are classified by searching the most similar images among the positive samples from */pos/index* and the negative samples from */neg/index* using the global image feature *JCD*. In addition to the on-screen output, 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 indices, but all threads share the same queue for the input data to be classified. Therefore, every image or video frame is only loaded once, 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, 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. 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 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. DATA AND DEMO

To show how the system performs, we used the ASU-Mayo Clinic polyp database [5]. It is at the moment 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 polyps in the dataset are diverse and vary in terms of shape, color and texture. 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. Table 1 gives an overview of all results.

| *Evaluation method* | *Precision* | *Recall* | *F1 score* |
|---|---|---|---|
| EIR - standard | 0.903 | 0.919 | 0.910 |
| MCB - standard | 0.683 | 0.683 | 0.683 |
| EIR - LOOC | 0.895 | 0.903 | 0.899 |
| MCB - LOOC | 0.636 | 0.636 | 0.636 |

**Table 1: Evaluation results of the detection subsystem. The table shows the EIR system and majority class baseline (MCB) performance for the standard train test set split evaluation and LOOC evaluation.**

First, we split the dataset into test and training sets. The test set contains two separate videos that are not used in the training dataset. To measure the performance, we used precision, recall and F1 score. 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, and an F1 score of 0.910. Remember that the best existing system, Polyp-Alert [13], achieved around 0.97 for the recall, but it was tested on another dataset. For these results, we used a fusion of the features JCD and OpponentHistogram, which we found to perform best in some additional experiments. The number of visual neighbours (size of the rank list returned by the search part of the classifier) was 71. The majority class baseline (MCB, all negative) is $0,683$ for precision, $0,683$ for recall and F1 score of $0,683$.

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 except 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 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 and the average F1 score is $0,899$. In comparison, the LOOC for the majority class baseline (all negative) has a precision of 0.636, a recall of 0.636 and a F1 score of 0.636. It is important to point out that we chose 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 (it ranges from 10 to 30 milliseconds depending of features used and resolution of the video). 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 parallization is not yet optimized and we are working on an even faster system, but this is out of scope for this paper.

Our system can also achieve higher a recall, at the cost

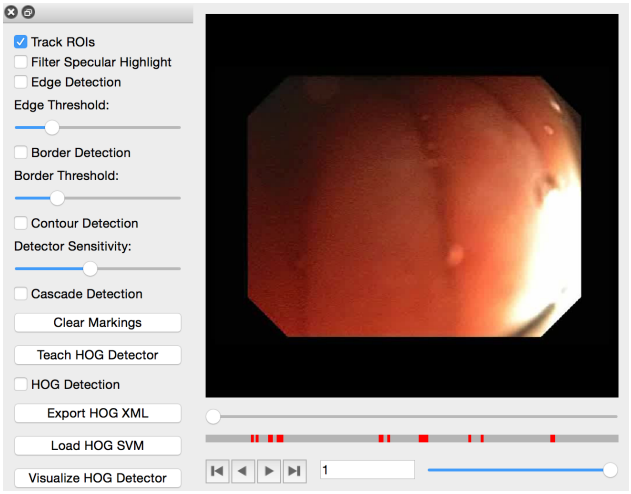



**Figure 3: Visualisation of the output. A positive finding is marked red on the timeline of the video.**

of the precision (or vice versa). For example, we can easily increase the recall by using more visual neighbours. 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 frames.

One may criticize us for using only polyp detection at the moment, but the Mayo data set is currently the only medical data set for our use case that is big enough and publicly available to show our performance. The system can easily be extend to different diseases by simply using a separate classifier for each category, which will make it better parallelizable and more accurate (since it is late fusion and late fusion has been proven as being more accurate [3]).

Possible ways to use the output of the classification tool are presented in the following figures. 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 analyse the whole video. Figure 3 shows the classification performance and how it is presented in a computer-aided diagnosis (CAD) tool for a standard colonoscopy video. One can see that the tool is able to classify videos in a way that can help experts to find irregularities but also help them reduce the time spent on video analysis.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we have presented an application for computer-aided diagnosis that can support medical doctors in analysing colonoscopy videos. We showed that we can reach high performance in terms of processing time, which would make it possible to use the system during live colonoscopies. At the same time, we reach high detection performance.

While extending the application to support multiple disease detection is trivial by adding more classifiers, the increased workload will also increase the total runtime of the detection algorithm. We strongly believe that if our tool is to be widely deployed and used by medical staff, it must be able to do classification and detection preferably during ongoing medical examinations, not only in post-examination analysis. A candidate for future improvement is therefore to run multiple classifiers of different diseases, like explored by Riegler et al. [10], in parallel by offloading the processing to multiple machines connected in a PCI Express network [8, 2]. This optimized version of the application will be able to dynamically allocate and release compute resources on demand from a pool of available GPUs. The use of multiple GPUs will also enable the system to run in real-time [7].

## 6. ACKNOWLEDGMENT

This work has been performed in the context of the FRINATEK project *EONS* (#231687) and the BIA project *PCIe* (#235530) funded by the Norwegian Research Council. The authors also acknowledge Zeno Albisser for his contributions.

## 7. REFERENCES

[1] 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*. ACM, 2015.

[2] L. B Kristiansen, J. Markussen, H. Kvale Stensland, M. Riegler, H. Kohmann, F. Seifert, R. Nordstrøm, C. Griwodz, and P. Halvorsen. Device lending in PCI express networks. In *Proc. of NOSSDAV*. ACM, 2016.

[3] H. J. Escalante, C. A. Hérnadez, L. E. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proc. of ACM ICMR*, pages 172–179. ACM, 2008.

[4] M. Lux. Content based image retrieval with lire. In *Proc. of ACM MM*, pages 735–738. ACM, 2011.

[5] Mayo-Clinic. Polyp dataset. http://polyp.grand-challenge.org/site/Polyp/AsuMayo/. [last visited, 04.04, 2016].

[6] 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.

[7] 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.

[8] 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 GPU. In *Proc. of MMSys*. ACM, 2016.

[9] 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.

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

[11] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Computer-aided detection of retroflexion in colonoscopy. In *Proc. of CBMS*, pages 1–6. IEEE, 2011.

[12] 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.

[13] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. de Groen. Polyp-alert: Near real-time feedback during colonoscopy. *CMPBM*, 2015.

[14] 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 BME*, 57(3):685–695, 2010.

# Paper XIV

# 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

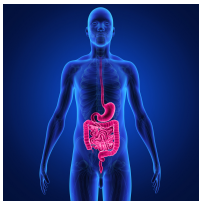


**Figure 1: GI tract (shutterstock.com)**

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).**

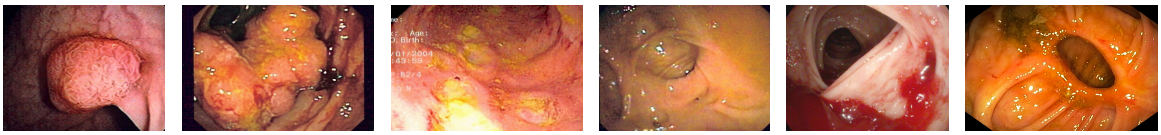
should raise concern from those that are better ignored. Obviously, live assistance has in the past been inhibited by excessive hardware costs, which prevented the creation and deployment of system that could perform in real-time. Our experimental prototype described in Section 4 makes use of modern 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 $256 \times 256$ to $400 \times 600$. 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 ("–").**

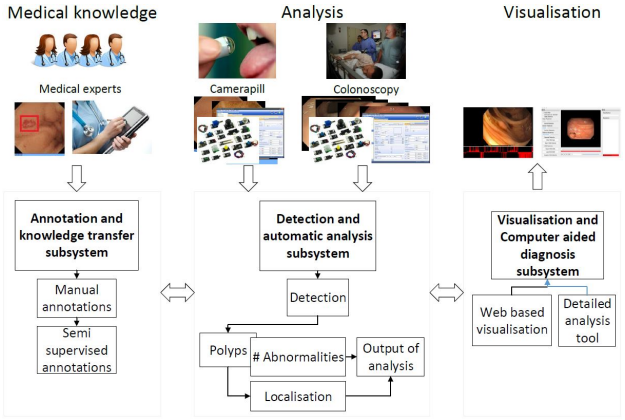



**Figure 4: EIR system: annotation and knowledge transfer, detection and automatic analysis and computer aided diagnosis.**

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 $1920 \times 1080$. 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 $856 \times 480$ resolution) and 453 (for $712 \times 480$ 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 $1920 \times 1080$, $856 \times 480$ and $712 \times 480$ 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

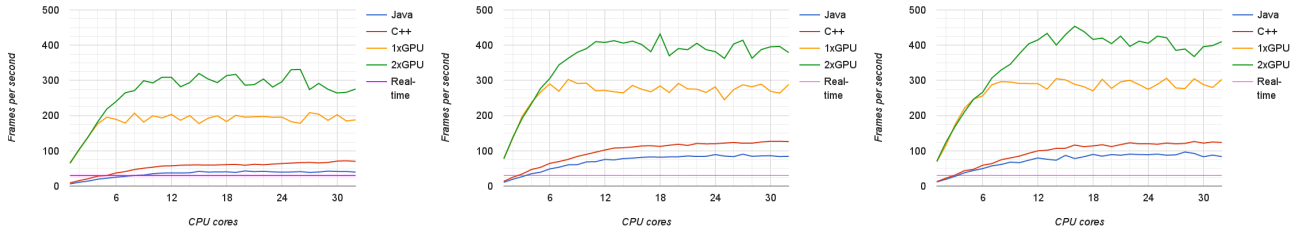
With 2.8 million cancer cases diagnosed in the GI system per year with a mortality rate of about 65%, we have the best motivation to perform research in the proposed area. The GI example that we used in this paper is only the tip of the iceberg of unsolved problems in the health care sector. By exposing 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 $1920 \times 1080$. (b) Videos with a resolution of $856 \times 480$. (c) Videos with a resolution of $712 \times 480$.

**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 $1920 \times 1080$, $856 \times 480$ and $712 \times 480$ 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 XV

# GPU-accelerated Real-time Gastrointestinal Diseases Detection

2016 IEEE 29th International Symposium on Computer-Based Medical Systems

# 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

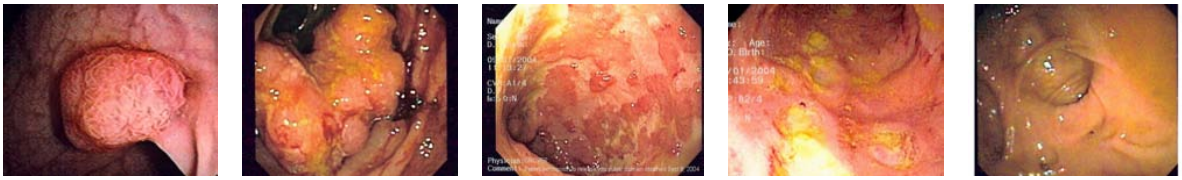
With the rapid developments in technology that allow miniaturization of cameras and sensors for moving them through the human body, there is an increasing need for real-time medical systems. These improvements lead to a lot of advantages for both patients and doctors, but also challenges for the computer 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.

2372-9198/16 $31.00 © 2016 IEEE
DOI 10.1109/CBMS.2016.63



(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

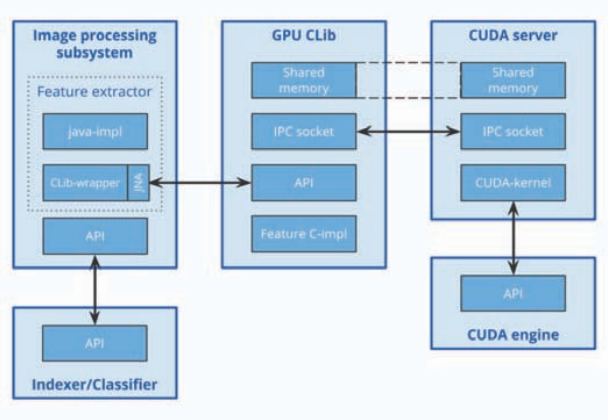



Fig. 3. The main processing application consisting of the indexing and classification parts uses the GPU-accelerated image processing subsystem. This subsystem provides feature extraction and image filtering algorithms. The most compute-intensive procedures are executed on a stand-alone CUDA-enabled 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 $1920 \times 1080$ pixels (full high definition) [16]. Further, we concentrate the experiment on the detection part.

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

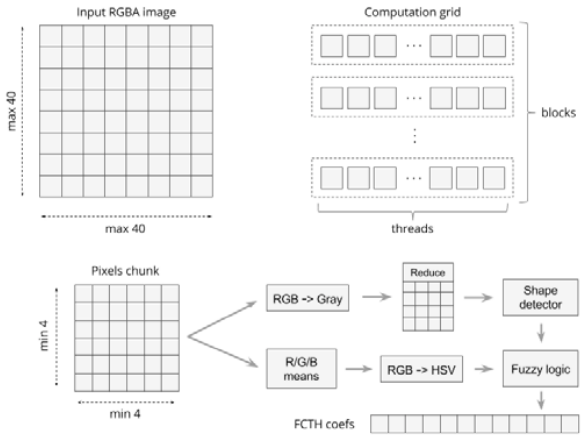



Fig. 4. GPU-acceleration is used to extract various features from input frames. The figure shows an example of our FCTH feature implementation. The input frame is split into a number of non-overlapping blocks. Each of them is processed separately by two GPU-threads. The main processing steps include 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 $856\times480$ and np_9 with $712\times480$. 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 ($1920 \times 1080$), WVGA1 ($856 \times 480$), WVGA2 ($712 \times 480$) and CIF ($384 \times 288$). 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

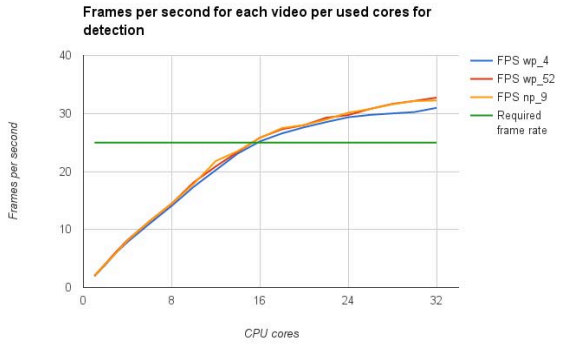



Fig. 5. The detection performs efficiently and the required frame rate is reached with 12 GB of memory and 16 CPU cores used in parallel on cluster-based computation platform without utilizing heterogeneous architecture.

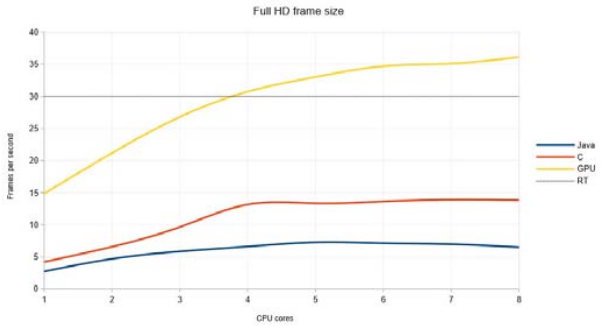



Fig. 6. The improved GPU-enabled heterogeneous algorithm reaches real-time performance (RT line) with 30 frames per second for full HD ($1920 \times 1080$) videos on a desktop PC using only 4 CPU cores and 5 Gb of memory. The maximum frame rate is around 36 FPS using 8 CPU cores. The Java and C implementations 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

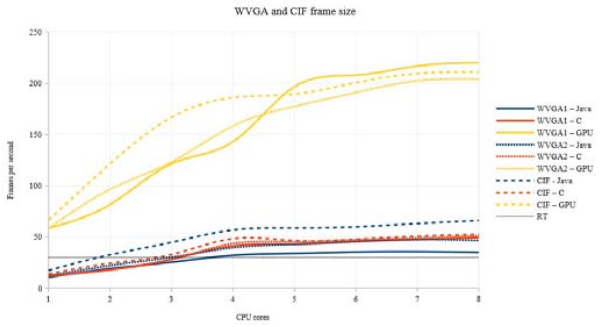



Fig. 7. The smaller WVGA1 ($856 \times 480$), WVGA2 ($712 \times 480$) and CIF ($384 \times 288$) videos can be processed by the improved GPU-enabled heterogeneous algorithm in real-time using only 1 CPU core. The maximum frame processing rate reaches more than 200 FPS. These results can be improved 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

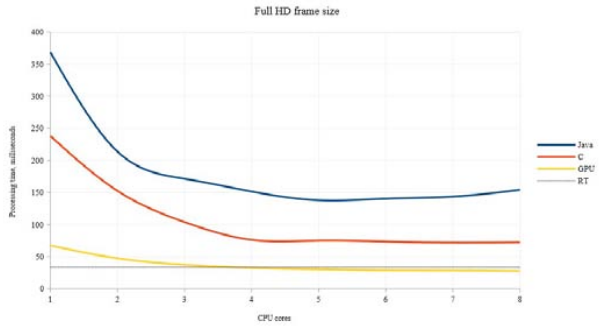



Fig. 8. The processing time for the GPU-accelerated algorithm decreases slightly with increasing number of used CPU cores for a single full HD frame. This happens due to the CPU-parallel implementation of feature comparison and search algorithms which are not as compute intensive as feature extraction. 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.

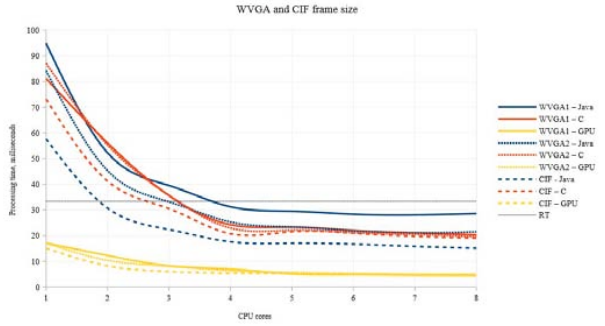



Fig. 9. For the smaller frame sizes the GPU-accelerated algorithm results in a processing time far below the real-time margin. The minimum is reached with 5 milliseconds using 8 CPU cores. This is a prove for the high system performance and ability to be extended by additional features or to process 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 XVI

# Device Lending in PCI Express Networks

# Device Lending in PCI Express Networks

Lars Bjørlykke Kristiansen[1], Jonas Markussen[2], Håkon Kvale Stensland[2], Michael Riegler[2], Hugo Kohmann[1], Friedrich Seifert[1], Roy Nordstrøm[1], Carsten Griwodz[2], Pål Halvorsen[2]

[1]Dolphin Interconnect Solutions AS, Norway
[2]Simula Research Laboratory, Norway & University of Oslo, Norway

{larsk, hugo, sfr, royn}@dolphinics.no
{jonassm, haakonks, michael, griff, paalh}@simula.no

## ABSTRACT

The challenge of scaling IO performance of multimedia systems to demands of their users has attracted much research. A lot of effort has gone into development of distributed systems that add little latency and computing overhead. For machines in PCI Express (PCIe) clusters, we propose Device Lending as a novel solution which works at a system level.

Device Lending achieves low latency and extremely low computing overhead without requiring *any* application-specific distribution mechanisms. For applications, the remote IO resource appears local. In fact, even the drivers of the operating system remain unaware that hardware resources are located in remote machines.

By enabling machines in a PCIe cluster to lend a wide variety of hardware, cluster machines can get temporary access to a pool of IO resources. Network cards, FPGAs, SSDs, and even GPUs can easily be shared among computers. Our proposed solution, Device Lending, works transparently without requiring any modifications to drivers, operating systems or software applications.

## CCS Concepts

**•Computer systems organization → Distributed architectures; •Software and its engineering** → *Distributed systems organizing principles;*

## Keywords

Multimedia, GPU, PCIe, interconnect, device sharing

## 1. INTRODUCTION

Performing multimedia tasks in real time are challenging and frequently require distributed systems. Tetzlaff et al. [28] early provided a classification for designing a distributed system. Actual implementations have often addressed requirements for low latency and high throughput by specialized interconnect networks [8, 6, 10, 7]. The PCI Express (PCIe) interconnect network [5, 19], which today is the dominant interconnection technology inside individual computers, can be connected to the internal networks of remote computers by using PCIe non-transparent bridges (NTB) [23]. The communication over such an interconnect network may be performed just like in classical interconnected networks, for example by implementing a high-performance TCP/IP stack for PCIe [11].

From the point of view of each computer, an NTB is just another PCIe device that offers memory areas for mapping into the remote computer's physical address space. An unusual property of the NTB, is that this memory is not located on it, but is rather a mapping of arbitrary memory areas within the domain of other computers that are also connected to the same NTB.

This raises the question whether all PCIe devices that are connected to any of the computers attached to such an NTB, can be considered part of one common resource pool. With Device Lending, devices can by lent by one computer into another without involving the CPU in data path forwarding.

All resources of any PCIe device are represented by mapped addresses, including their control registers and interrupts, so all of them can be mapped by an NTB. Obviously, such mapping cannot be trivial. Whereas data areas can be mapped into a computer's address space just like those of locally installed devices, a reverse mapping is required for interrupts. Furthermore, devices can be lent dynamically by one computer to another only if the operating systems can handle that PCIe devices are added to and removed from their address space, i.e., if they have hotplug support [14] for the specific device.

Once these problems are solved, we can see that the power of this approach goes far beyond the classic interconnection challenges of a streaming server. Within a small cluster, devices can be pooled together and time-shared by different

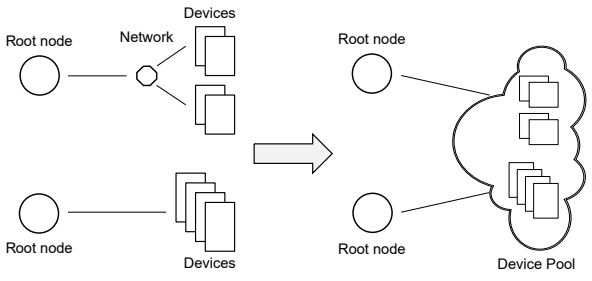



Figure 1: PCIe devices on separate machines could be pooled together and shared between multiple computers.



*NOSSDAV'16, May 13 2016, Klagenfurt, Austria*



ISBN 978-1-4503-4356-5/16/05…$15.00

DOI: http://dx.doi.org/10.1145/2910642.2910650

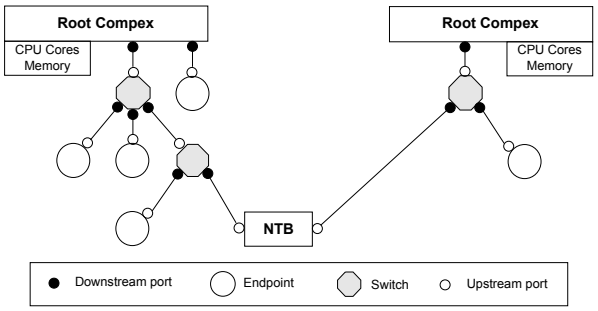



Figure 2: An example of a PCIe topology.

computers (Figure 1). Network cards can be assigned to a computer while it needs high throughput. Instead of copying data between SSD disks over traditional network, the disk can be borrowed and accessed directly. For a large CUDA programming task, a computer can lend additional cards and use CUDA's own peer-to-peer model instead of relying on additional middleware like rCUDA [4]. Pogorelov et al.[21] have shown how a multimedia workload can be offloaded to a remote GPU using Device Lending.

In this paper, we present how we achieve this pooling of PCIe devices using only native device drivers. We present the state of our proof-of-concept implementation of Device Lending for Ethernet network cards and SSD disks, and in more detail, our prototype for GPU lending. We show that the GPUs can be lent dynamically without any modifications to drivers or user-space applications.

The paper is organized as follows: we present essential capabilities of PCIe in Section 2. Section 3 addresses the current state of PCIe virtualization support. In Section 4 we discuss related work. Section 5 goes into details of our implementation of Device Lending, followed by performance results for GPU lending in Section 6. Conclusion and further opportunities are discussed in Section 7.

# 2. PCI EXPRESS

PCIe is an industry standard for architecture-independent connection of hardware peripherals to computers. In PCIe terminology, such a peripheral is a *PCIe endpoint*. While its predecessor PCI relied on parallel buses that were shared between endpoints, PCIe uses point-to-point links (still called *buses*) that consist of 1 to 32 *lanes*. These buses can be connected to *PCIe switches*, which may be connected to other switches, forming a tree structure where endpoints are leaves, switches are inner nodes, and buses are edges. An example of a PCIe topology is illustrated in Figure 2. The connection of a bus to a switch is called a *port*, but (primarily to illustrate how backwards compatibility with PCI is achieved) it is also known as a *bridge*. Ports towards the tree root are called *upstream*, the other *downstream*. The network of buses, endpoints and switches is referred to as *fabric*. For communication, PCIe specifies a layered protocol structure, whose upper layer is called transaction layer, exchanging transaction layer packets (TLPs). Routing occurs in a strictly hierarchical fashion, i.e., packets do not need to pass through the root of the tree.

At the root of the PCIe tree is the *root complex*, which an implementation can either interpret as an endpoint that is connected to the root node of the fabric or as being the root node. In this paper, we refer to the root complex as the root node. Directly connected to the root complex is the CPU core and memory controller. Each endpoint may act like a group of distinct devices. Each of these is called a *function* and is separately addressable by the triplet of its *bus*, *device* and *function* IDs, referred to as its *BDF*.

Both endpoints and buses are detected by reading their *configuration space*. At system boot, the *system* (BIOS or OS) scans possible BDFs for vendor IDs in a process called bus enumeration. If an endpoint or bus is present at a given BDF, the system reads the associated configuration space. This contains data structures in a standardized format [19], allowing the device to define its requirements.

## 2.1 Memory-mapped IO

When a configuration space is found at a given BDF, the system reads the its Base Address Registers (BARs) to determine the function's size requirements and number of address spaces that must be mapped into the host's linear address space. This mapping allows the CPU to access device registers of the endpoint through regular memory accesses. This process is called Memory Mapped IO (MMIO) and allows memory operations to be transparently translated into TLPs by devices and the CPU.

The system writes the mapped addresses into the BARs, which allows the endpoint to interact with the host machine. If the device has an onboard Direct Memory Access (DMA) engine, it can be instructed to read from and write to any memory buffers directly, including main memory and other endpoints. Without a DMA engine, the CPU must write to MMIO registers to transfer data.

### 2.1.1 Posted and non-posted transactions

Some PCIe requests require end-to-end notification upon completion. These requests are called *non-posted transactions*, while requests that do not require notification are *posted transactions*. A memory write request is an example of a posted transaction. The requester sends the write request along with the data and after it leaves the egress port it is no longer the responsibility of the requester. Memory read requests, on the other hand, requires explicit completion TLPs.

Non-posted requests are significantly affected by the length of a PCIe path. The longer the path, the higher the request-completion latency becomes. In addition, the number of read requests in flight is limited by how many the requester supports. The number of supported read requests in flight has an impact on read performance.

### 2.1.2 Transparent bridges

A switch is associated with one contiguous address range in the host address space and is aware of it. The address range is called *address window*, and spans all address ranges assigned to endpoints downstream of this switch. Each port on the root complex is associated with its own contiguous address range. This allows shortest-path routing in the tree based on physical address. Switches and their ports perform only routing in this scenario, and are *transparent* in that sense. PCIe bridges can be regarded as transparent bridges.

### 2.1.3 Non-transparent bridges

It is desirable to extend PCIe out of the single computer and use it for high-speed interconnection networks due to its high bandwidth and low latency [22]. One way of doing this

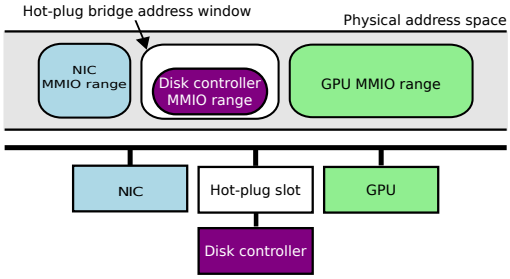



Figure 3: Physical address ranges are reserved by OS or BIOS at boot time. Memory requirements of hot-plugged devices must fit within the already existing address windows.

is by using NTBs [23]. Although not standardized, NTBs are widely adopted and all NTB implementations have similar capabilities. Several processor architectures, including recent Intel Xeon CPUs, support NTB implementations [26].

Despite the name, NTBs do actually appear as PCIe endpoints in one or more PCIe fabrics at the same time. They are mapped with large MMIO areas similar to other endpoints. However, unlike other endpoints and like transparent bridges, memory operations on these areas are forwarded from one fabric into another. Since an NTB is mapped differently in each host's address space, it performs address translation on the TLPs during forwarding. This address translation is similar to a single-level page table. Effectively, NTBs create a shared memory architecture across several hosts [13].

However, an NTB address space is not necessarily linear. Its MMIO area is divided into equally sized segments, and each segment can be mapped anywhere into the remote host's address space. This is done by replacing part of the address with a per-segment offset into the remote host's address space. Not only does this allow a remote host to access local RAM memory, it also enables a remote host to access MMIO areas of local PCIe devices.

## 2.2 Message-signaled interrupts

Whereas physical interrupts lines were used in traditional PCI, PCIe uses *Message-Signalled Interrupts* (MSI) [17, 19]. When an endpoint issues an MSI, this is actually a normal memory write to a special address, which is then interpreted by the chipset and used to generate an interrupt to the CPU. For our work, this has the essential implication that the address of an MSI can be mapped through an NTB.

## 2.3 Hot-plugging

The idea of lending devices without any OS changes whatsoever includes the goal that the devices must appear to and disappear from the OS at run-time. Obviously, there are device drivers that are not capable of coping with run-time appearance or disappearance. We can address the challenges that occur on a level "underneath" the OS.

PCIe specifies the ability of hot-plugging devices, making them available to the system while it is running. This ability was designed for replacing devices without rebooting the machine [22, 14]. Consequently, most OS implementations reserve MMIO ranges at boot time and keep them unchanged until reboot.

This is sufficient for hot-plugging in the sense of *hot-replace*, but problematic for *hot-add*, as shown in Figure 3.

When a device is hot-plugged, it appears in a port of a PCIe switch whose contiguous address range has already been mapped. A worst-case reservation for an arbitrary endpoint for every hot-plug capable port of a switch is not usual but may be feasible. However, a hot-add operation may plug an entire subtree of devices into the port, with an arbitrarily large requirement for MMIO range. If the required address range is too large, a remapping of the host address space must be undertaken. This is, however, non-trivial, and few OSes support it currently. In our implementation, the hot-add variant of hot-plugging becomes trivial, as devices become accessible through the NTB. The already allocated address space is large enough to contain all the MMIO areas.

# 3. VIRTUALIZATION SUPPORT IN PCIE

Traditionally, virtualization has been used to provide host resources to guest OSes in virtual machines (VM). Since endpoints are already mapped into the host address space, and the VM has a different memory layout than the host, they can traditionally not access endpoints without specialized drivers in the guest OS, which are aware of the mapping. Due to the performance penalty of this (and the breach of VM isolation that a common memory layout would bring), dedicated virtualization units have been introduced.

## 3.1 IO Memory Management Unit

By organizing memory in pages and adding a software-defined page-table, a Memory Management Unit (MMU) can translate addresses accessed by the CPU before passing them to chipset and memory controller. The MMU provides every processes in the host OS as well as every guest OS in a VM their own virtual, linear address space, while the physical memory can be fragmented or non-existent (e.g., swapped out).

The IO Memory Management Unit (IOMMU) [9] is similar to an MMU, but it provides virtualization of addresses between chipset (including CPU cores and MMU) and PCIe fabric. One of the most important features of the IOMMU is the DMA remapper, which translates addresses of memory operations from any IO device. In other words, it translates IO virtual addresses to physical addresses.

Similarly to pages mapped by an MMU, an IOMMU can group PCIe functions into *domains*, where each domain has separate mappings and its own address space. Such a domain can be part of the address space of a VM, while other PCIe functions remain isolated from the VM. This allows the VM to interact directly with the device using native device drivers in the guest OS, often referred to as *PCIe passthrough*.

Importantly, there is nothing that prevents the IOMMU from performing such a mapping for the host OS as well. This is an opportunity for Device Lending.

## 3.2 Single-Root IO Virtualization

Unlike the MMU's page maps, IOMMU mappings are not process-specific. Since IOMMU supports only one mapping per PCIe function, it can only assign an endpoint function to a single VM at a time. Single-Root IO Virtualisation (SR-IOV) [20] addresses this. SR-IOV-aware device can allow single physical PCIe functions to act as multiple virtual PCIe functions, allowing SR-IOV to map a single physical function to several VMs.

### 3.3 Performance penalty

As with most abstractions, DMA remapping brings a performance overhead. The translation tables are held in memory like the MMU's. When a memory access passes through, the IOMMU must perform a multi-level table look-up. Furthermore, it is located in the root complex, and all TLPs must be routed through the root to perform DMA remapping. In addition, unpredictable access patterns using small-sized pages can lead to thrashing of the IO translation look-aside buffer. PCI-SIG has developed an extension of the transaction layer protocol that allows caching of mapped addresses on the PCIe devices [19], but this is not widely available yet.

## 4. RELATED WORK

The idea of a unified bus for the inner components of a computer with those of another is not new. It was imagined for both ATM [24] and SCI [1]. These ideas never got implemented, because none of these technologies were picked up for the internal interconnection networks of computers.

PCIe is the dominant standard for the internal interconnection network. It is also proving to be a relevant contender for an external interconnection network. PCIe, however, was designed to be used within a single computer system only. In this section, we will discuss some solutions for sharing IO devices between multiple hosts.

### 4.1 Alternative protocols

There are several interconnection technologies, which are more widely adopted for creating high-speed interconnection networks than PCIe. These include InfiniBand, as well as 10Gb Ethernet. They may achieve the same throughput on interconnection links, but they are not integrated as closely with the system fabric as PCIe, and require soft-processing of protocol stacks. Their latency is therefore, inevitably, higher than that of PCIe interconnects.

### 4.2 Multi-Root IO Virtualization

Multi-Root IO Virtualization (MR-IOV) [18] specifies how several hosts can be connected to the same PCIe fabric. The fabric is logically partitioned into separate virtual hierarchies, where each host sees its own hierarchy without knowing about MR-IOV. MR-IOV require multi-root aware PCIe switches, and, in the same way as SR-IOVs require SR-IOV-aware devices to provide functions to several VMs, devices must be multi-root aware to provide functions to several virtual hierarchies (and thus hosts) at the same time.

Despite being standardized in 2008 [18], we are not aware of any MR-IOV-capable devices and very few switches. Instead, there are attempts to achieve MR-IOV-like functionality through a combination of SR-IOV with NTB-like hardware [27].

### 4.3 Ladon and Marlin

Our Device Lending idea is apparently timely, because very similar functionality was proposed in Cheng-Chun Tu et al. in the form of the Ladon [29] and Marlin [30] systems.

Ladon uses all PCIe and virtualization features as proposed in this paper, but it achieves less freedom than our Device Lending. In Ladon, PCIe devices that are offered for sharing are all managed by a dedicated computer, the management host. The only task of the management host is to manage sharing of the devices. The guest OSes that include these devices into their PCIe fabric are, first, all running in VMs, and second, they include the remote PCIe devices in their fabric for the entire lifetime of the OS. With our Device Lending, we can actually pool the resources of a small cluster of NTB-connected devices by lending in arbitrary direction. We can even exchange devices, and do this under the control of a running OS, not a dedicated machine. By combining PCIe hot-plug support in the OS with use of the NTB, we can insert remote PCIe devices while the OS is running. Finally, for devices whose native device drivers support hot-remove, we can stop borrowing without rebooting.

Marlin [30] can share network IO capacity in a cluster by forwarding Ethernet packets underneath the host's TCP/IP stack to another node, using an Ethernet-over-PCIe driver for legacy software and a dedicated stack for zero-copy mode. While this replicates Dolphin Interconnect Solutions' (Dolphin) SuperSocket approach [12], which is a continuation of SuperSockets for SCI [25], the technique appears generic for all interconnection technologies. With Device Lending, however, we borrow the network card from the remote host and require neither driver nor encapsulation overhead.

## 5. IMPLEMENTATION

We have implemented Device Lending for an unmodified Linux kernel, using an NTB and the IOMMU. The implementation is composed of two parts, the *lending* side and the *borrowing* side. For our proof-of-concept implementation, we rely on a NTB implementation from Dolphin, namely the PXH810 host adapter [2].

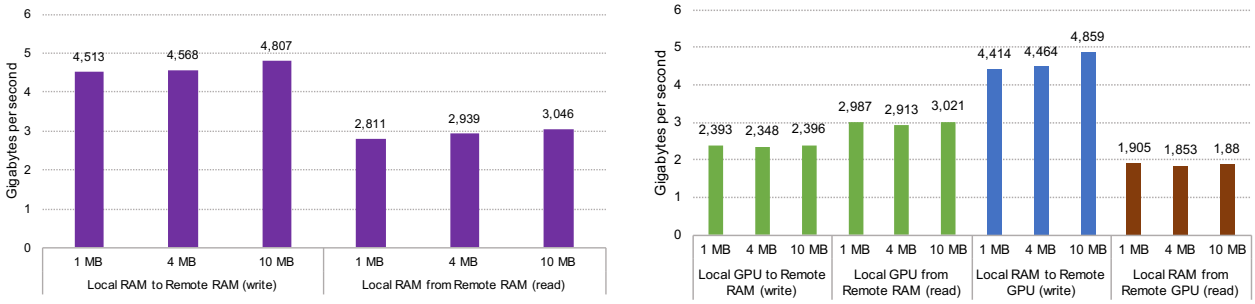
The lending side kernel module binds itself as a driver for the targeted PCIe devices. This provides us with exclusive access to the device, allowing the kernel module to access the device's configuration space while preventing other drivers on the host from interfering. The kernel module then notifies the borrowing side of all available devices.

When the user requests an available device, the borrowing side kernel module communicates with the lending side kernel module in order to read the device's configuration space. The lending side sets the targeted device into a per-borrower IOMMU domain, isolating the device from the rest of the system and other devices. The borrowing side then sets up the necessary MMIO mappings using the NTB and tells the lending side to set up the reverse mappings for device to RAM DMA as well as MSI mappings. Following this, the borrowing side then injects the device into the Linux PCI subsystem and signals a hot-add event. Linux will probe the device, set it up and load the device driver.

The device driver is now able to communicate with the device using MMIO access. Whenever the device driver sets up new DMA mappings using the Linux DMA-API, the borrowing side kernel module intercepts these calls and dynamically sets up and tear down the necessary IOMMU mappings. This allows the borrowing side device driver to transfer data to the remote device with no additional software overhead.

## 6. EVALUATION AND DISCUSSION

As the global address space feature of PCIe is unique, and since, to the best of our knowledge, no MR-IOV implementations exist, our Device Lending concept has few relevant




(a) RAM to RAM

(b) RAM to GPU

Figure 4: DMA transfer bandwidth across the NTB with different transfer sizes. The DMA engine on the NTB is used.

comparisons. Alternative solutions either require extensive virtualization support or additional protocol stacks. In order to evaluate our proof-of-concept implementation, we therefore evaluate the performance compared to what is possible to achieve with specialized use of the NTB. To establish a point of reference, we measured RAM to RAM bandwidth as this shows the maximum possible transfer rate.

We configured two test machines, shown in Figure 5. Both machines have a single Nvidia Tesla K40 directly connected to the root complex each. The machines were connected together using two x8 Gen3 Dolphin PXH810 adapter cards and an external PCIe cable. In all our tests, Machine A was used to initiate transfers.

## 6.1 Reference evaluation

For our RAM to RAM reference, we transferred data between the two machines over the NTB and measured the bandwidth without Device Lending (Figure 4). Here, we used Dolphin's SISCI API for programming the DMA engine on the NTB itself [3, 16]. All PCIe endpoints in our setup are connected directly to the root complex, which is why transferring between remote RAM and local RAM shows the optimal performance over the NTB (Figure 4a). RAM to remote RAM latency is approximately 573 ns.

Write requests peak at around 4.8 GB/s on our test configuration, shown on the left-hand side in Figure 4a. As mentioned in Section 2.1.1, memory read requests are affected by the distance in the PCIe hierarchy because they are *non-posted* transactions. However, there are is an additional factor that also limit the performance of read operations. PCIe defines a *maximum read request size*. This is configured by the system to ensure that the bandwidth is shared among all the devices in the hierarchy. For our test system, the maximum read request size is 512 bytes, and the TLP maximum payload size is 128 bytes. The DMA engine on the NTB handles 64 read requests in flight. As seen in Figure 4a, read requests peak at around 3 GB/s on our configuration.

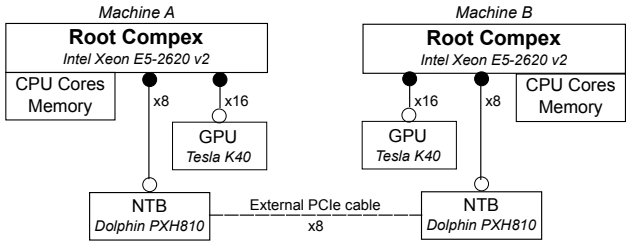



Figure 5: The setup used for our evaluation

Since the GPU is even further away than RAM, as illustrated in Figure 5, we see a considerably lower bandwidth for RAM to remote GPU and GPU to remote RAM transfers. Figure 4b shows the results of using the DMA engine on the NTB. The two scenarios on the left-hand side show using a local GPU on Machine A and RAM on Machine B. The two other scenarios on the right-hand side show the opposite, using local RAM on Machine A and the remote GPU on Machine B. It is important to note that when using a local GPU, the DMA engine on the NTB first has to perform read requests to the GPU before it is able to push it to the remote side using write requests. In other words, it is a two-part operation. It is interesting to note that reading from a local GPU and pushing it to remote RAM (Figure 4b, second from left) is is similar to reading from remote RAM (Figure 4a, on the right). This indicates that the latency added by the NTB is around the same as having to route TLPs through the root complex.

## 6.2 Device Lending evaluation

One of the novel properties of Device Lending is that it can be achieved with no modifications to endpoint devices or device drivers or even user-space software. We therefore wanted to use an already existing benchmarking tool. A well-known tool in the CUDA developer community, is the `bandwidthTest` [15] utility. This tool is included in the CUDA Toolkit samples. In default mode of operation, this program allocates page-locked buffers in RAM and measures the bandwidth it achieves when copying to the GPU and vice versa using the GPUs onboard DMA engine. We argue that making one of the most complex proprietary GPU drivers on the market work with our implementation serves as good test for our proof-of-concept.

In our setup, Machine B was configured to lend its Tesla K40 GPU to Machine A, making it available for the OS and driver on the remote machine. Figure 6 shows the results of running `bandwidthTest` on the remote Tesla K40 using different transfer sizes. The left side shows the results of making the onboard DMA engine write to remote RAM on Machine B (around 4.9 GB/s), while on the right we see the results of making the onboard DMA engine read data from remote RAM (around 2 GB/s). These numbers are comparable to the numbers seen in Figure 4b, as they show a similar scenario. However, as they use different DMA engines, they also have different locality to the data.

Using the onboard DMA engine to write to remote RAM is close to the speeds for local RAM to remote RAM trans-

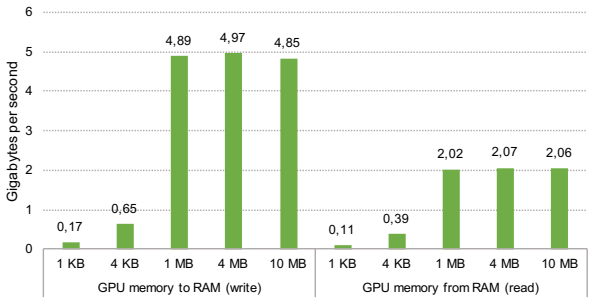



Figure 6: `bandwidthTest` running on a borrowed GPU. The DMA engine on the GPU is used to transfer.

fers (around 4.9 GB/s). Reading from remote RAM and pulling it to GPU memory (around 2 GB/s) is a bit slower than reading from remote RAM and writing it to local RAM (around 3 GB/s). This is caused by the onboard DMA engine on Machine A's GPU being even further away from the remote RAM on Machine B than the DMA engine on Machine A's NTB.

# 7. CONCLUSION AND FUTURE WORK

In this paper, we presented the Device Lending concept, which allows a cluster of PCIe-connected computers to establish a pool of PCIe devices. These devices can subsequently be time-shared in a process of lending and borrowing. Since these devices appear like hot-plugged local devices to the borrowing OS, even the host OS can use them with their native drivers. For all native device drivers that support hot-plugging, these borrowed devices can be returned without rebooting. Having built the infrastructure for this, we demonstrated its performance in this paper, and provide hints for the best possible use of borrowed devices.

In further work, we will investigate concurrency challenges when multiple devices are borrowed and situations where the lender needs to take the device back forcefully. We are also planning to implement a framework for managing Device Lending. In addition, we are investigating the possibility for lending separate functions of SR-IOV devices in order to implement MR-IOV without needing specialised hardware.

## Acknowledgments

This work has been performed mainly in the context of the BIA project *PCIe* (#235530) funded by the Research Council of Norway (RCN), with contributions from EONS (RCN #231687) and POPART (EU H2020 #644874). The authors also acknowledge Magma for providing Nvidia Tesla GPUs.

# 8. REFERENCES

[1] K. Alnæs, E. H. Kristiansen, D. B. Gustavson, and D. V. James. Scalable coherent interface. In *Proc. of CompEuro*, pages 446–453, 1990.

[2] Dolphin Interconnect Solutions. PXH810 Gen3 PCI Express NTB Host Adapter.

[3] Dolphin Interconnect Solutions. SISCI API.

[4] 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.

[5] T. Fountain, A. McCarthy, and F. Peng. PCI express: An overview of PCI express, cabled PCI express and PXI express. In *Proc. of ICALEPCS*, 2005.

[6] S. Ghandeharizadeh, R. Zimmermann, W. Shi, R. Rejaie, D. Ierardi, and T.-W. Li. Mitra: A scalable continuous media server. *Springer Multimedia Tools and Applications*, 5(1):79–108, 1997.

[7] R. S. Grover, Q. Li, and H.-P. Dommel. Performance study of data layout schemes for a SAN-based video server. *Parallel Computing*, 34(12):747–756, 2008.

[8] J. P. Hayes, T. N. Mudge, Q. F. Stout, S. Colley, and J. Palmer. Architecture of a hypercube supercomputer. In *Proc. of ICPP*, pages 653–660, 1986.

[9] Intel Corporation. *Intel Virtualization Technology for Directed I/O*, 2014.

[10] T. Jones, A. Koniges, and R. Yates. Performance of the IBM general parallel file system. In *Proc. of IPDPS*, pages 673–681, 2000.

[11] V. Krishnan. Evaluation of an Integrated PCI Express IO Expansion and Clustering Fabric. In *Proc. of HOTI*, pages 93–100, 2008.

[12] V. Krishnan, T. Comins, R. Stalzer, and D. Wong. A case study in I/O disaggregation using PCI express advanced switching interconnect (ASI). In *Proc. of HOTI*, pages 15–24, 2006.

[13] L. B. Kristiansen. PCIe Device Lending: Using Non-Transparent Bridges to Share Devices. Master's thesis, University of Oslo, 2015.

[14] G. Kroah-Hartman. How the PCI hot plug driver filesystem works. *The Linux Journal*, 97(2), May 2002.

[15] NVIDIA Corporation. *CUDA Toolkit Documentation 7.5*, 2015.

[16] NVIDIA Corporation. *GPUDirect Technology Overview*, 2015.

[17] PCI-SIG. *PCI Local Bus Specification*, 2002.

[18] PCI-SIG. *Multi-root I/O Virtualization and Sharing Specification*, 2008.

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

[20] PCI-SIG. *Single-root I/O Virtualization and Sharing Specification*, 2010.

[21] K. Pogorelov, M. Riegler, J. Markussen, M. Lux, H. K. Stensland, T. Lange, C. Griwodz, P. Halvorsen, D. Johansen, P. T Schmidt, and S. L. Eskeland. Efficient processing of videos in a multi auditory environment using Device Lending of GPUs. In *In Proc. of MMSys*. ACM, 2016.

[22] M. Ravindran. Extending Cabled PCI Express to Connect Devices with Independent PCI Domains. In *Proc. of IEEE Systems Conference*, pages 1–7, 2008.

[23] J. Regula. *Using Non-transparent Bridging in PCI Express Systems*. PLX Technology, Inc, 2004.

[24] K. Saito, K. Anai, K. Igarashi, T. Nishikawa, R. Himeno, and K. Yoguchi. ATM bus system. US 5,796,741 A, 1998. US patent.

[25] F. Seifert and H. Kohmann. SCI SOCKET - a fast socket implementation over SCI. 2006.

[26] M. J. Sullivan. Intel Xeon Processor C5500/C3500 Series Non-Transparent Bridge. Technical report, 2010.

[27] J. Suzuki, Y. Hidaka, J. Higuchi, T. Baba, N. Kami, and T. Yoshikawa. Multi-root Share of Single-Root I/O Virtualization (SR-IOV) Compliant PCI Express Device. In *Proc. of HOTI*, pages 25–31, 2010.

[28] W. Tetzlaff, M. Kienzle, and D. Sitaram. A methodology for evaluating storage systems in distributed and hierarchical video servers. In *Compcon Spring, Digest of Papers.*, pages 430–439, 1994.

[29] C.-C. Tu, C.-t. Lee, and T.-c. Chiueh. Secure I/O device sharing among virtual machines on multiple hosts. *SIGARCH Comp. Arch. News*, 41(3):108–119, 2013.

[30] C.-C. Tu, C.-t. Lee, and T.-c. Chiueh. Marlin: A memory-based rack area network. In *Proc. of ANCS*, pages 125–136, 2014.

# Paper XVII

# 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

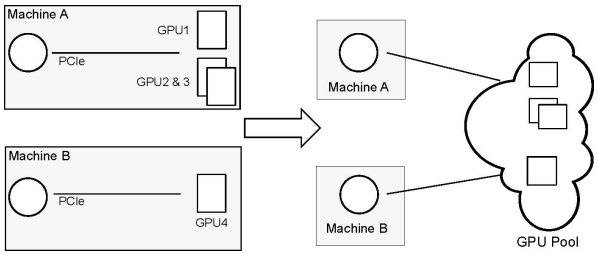



**Figure 2: Pooling of devices attached in the PCIe network in the experimental setup.**

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.**

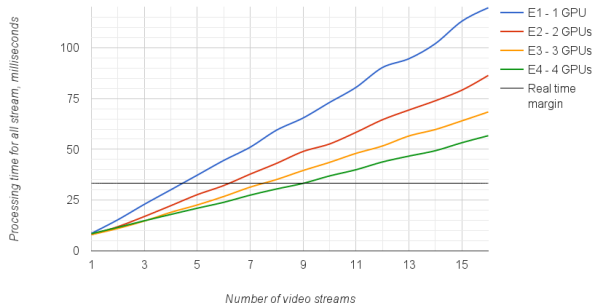
reveal that for up to 7 parallel full HD streams, the 3 local GPUs are fast enough. For more than 7 streams, GPU lending is required. The graph shows that the more parallel streams are processed, the better is the performance gain from the borrowed GPU. We assume that this is due to the excessive overhead 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.

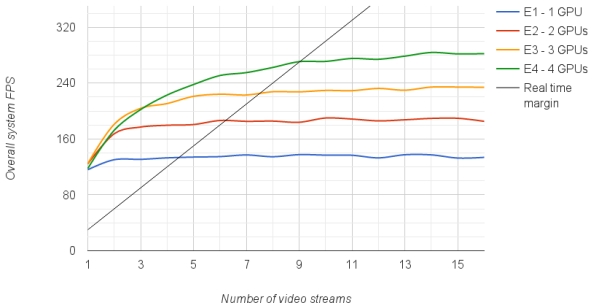
All in all, the experiments showed two important things: (i) Device Lending does not make sense for small amounts of data, but if the data to process is large it can give a large performance boost, and (ii) Device Lending makes sense in a multi-auditory scenario like we present with our demo.

# 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.**

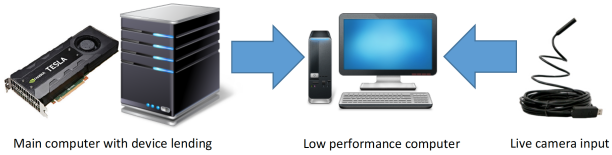



**Figure 4: A compete overview of the demo setup. The demo consists of 2 computers, 1 Dolphin interconnect device, 1 screen, an artificial colon and a flexible camera. The users can use the camera in the flexible colon and will get real-time feedback about possible findings. Furthermore, the demo can 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.

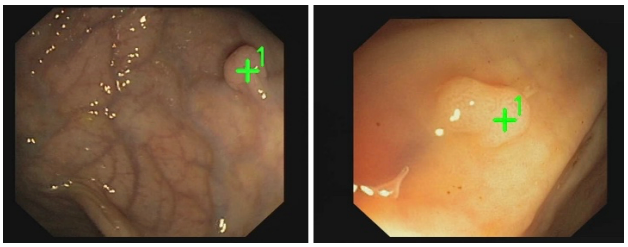


**Figure 5: This figure shows 2 examples of what the doctor will see on the screen and what we will show during the demo. In both pictures, the system detected polyps and marked them with a cross. If nothing is detected, the corners of the screen are marked green for feedback.**

## 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.