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

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.

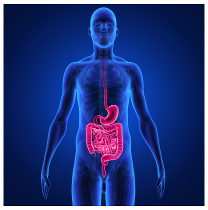


Figure 1: GI tract (shutterstock.com)

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 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.

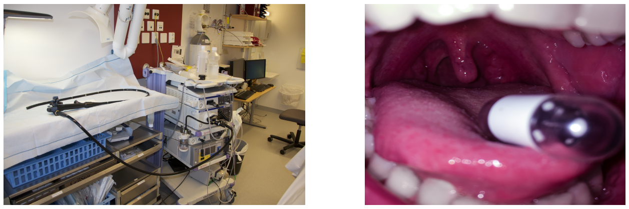

(a) Colonoscopy equipment (b) VCE capsule (camera pill)

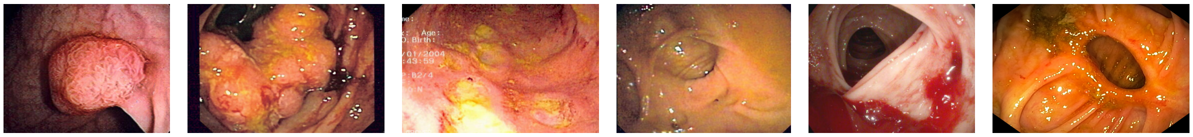
Figure 2: Endoscopy vs. wireless capsule endoscopy (VCE).

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

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

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

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



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

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

2. HEALTH MULTIMEDIA CHALLENGES

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

Nevertheless, there are unmet needs and potentials for improvements, which can be remedied by introducing better and more efficient digital medical systems. For colorectal cancer (CRC), which has one of the highest incidences and mortality of the diseases in the GI tract, early detection is essential for a good prognosis and treatment. Minimally invasive endoscopic and surgical treatment is most often curative in early stages (I-II) with a 5-year survival probability of more than 90%, but in advanced stages (III-IV), radiation and/or chemotherapy is often required, and it has a 5-year survival of only 10-30% [6].

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

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

3. RELATED WORK AND NEW TRENDS

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

3.1 GI Tract Endoscopy Technology

There are several providers of endoscopy systems and VCE devices. Last generation equipment for manual procedures like colonoscopy and gastroscopy provides video with high resolution and high frame rates. There is, however, no computer-aided diagnostic feedback. In this respect, Polyp-Alert [40] is the most promising with polyp detection capabilities, but with the main purpose of evaluating how well the procedures are performed. For live analysis of endoscopy videos, our target system aims to go far beyond the currently existing systems. The other approach to record videos of the GI tract is VCEs using a small capsule type device (a 11mm×25mm pill), which has at least one image sensor, antenna, battery, light source and wireless transceiver. The capsule is swallowed to record the GI tract. There are several vendors providing such capsules, like IntroMedic, CapsoVision, Medtronic (Given) and Olympus. The current VCEs often have a variable framerate (increasing the framerate to about 30-35 FPS when entering the small intestine), but a rather low resolution ranging from 256×256 to 400×600. One of the main challenges for use of VCEs is man-hours of medical staff required for analysis. There are about 216,000 images per examination, and a very experienced endoscopist needs at least 30 to 60 minutes to process the video and possible sensor data. Therefore, it is important to develop automatic methods that can reduce the burden on medical staff and speed up the analysis of the videos. Currently, the software can segment the videos and can allow endoscopists to fast forward and look at multiple videos at the same time (probably affecting the detection accuracy). Moreover, some software includes small detection components that provides only vague "hints", for example about the detection of the color red, which may indicate bleeding. Other main limitations with VCEs are that the lack of means for cleaning particles (food/stool) in the bowels, and their uncontrolled forward movement through the bowel that cannot be guided to take a close-up picture or a tissue sample from detected lesions.

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

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

3.2 Abnormality Detection

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

4. SHOWCASE FOR HOW-TO MULTIMEDIA IN MEDICINE

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

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 ("–").

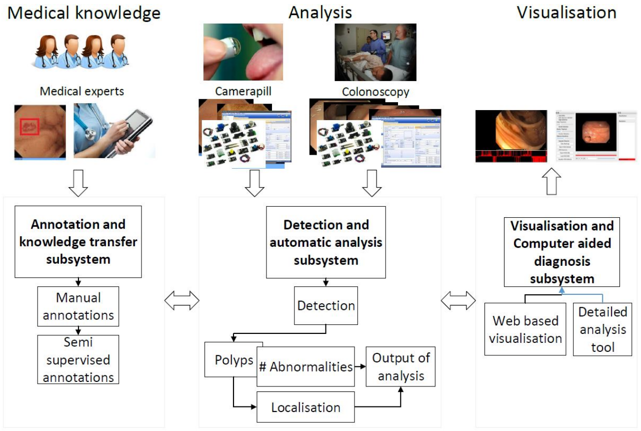


Figure 4: EIR system: annotation and knowledge transfer, detection and automatic analysis and computer aided diagnosis.

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

4.2 Detection and Automatic Analysis

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

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

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

4.3 Visualization and Diagnosis

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

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

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

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

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

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

5. PRELIMINARY RESULTS

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

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

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

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

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

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

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

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

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

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

6. OUTLOOK AND CHALLENGES

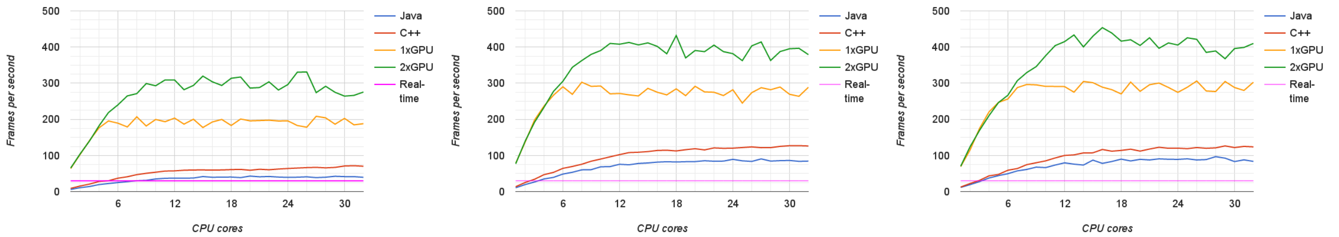
With 2.8 million cancer cases diagnosed in the GI system per year with a mortality rate of about 65%, we have the best motivation to perform research in the proposed area. The GI example that we used in this paper is only the tip of the iceberg of unsolved problems in the health care sector. By exposing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

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

6.1 Open Challenges

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

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



(a) Videos with a resolution of 1920×1080.

(b) Videos with a resolution of 856×480.

(c) Videos with a resolution of 712×480.

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

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

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

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

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

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

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

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