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

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

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

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

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

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

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

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

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

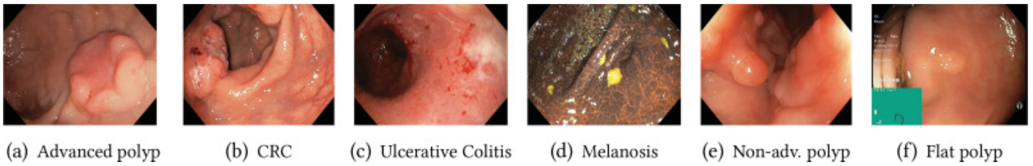
1. INTRODUCTION

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

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

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

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



(a) Advanced polyp (b) CRC (c) Ulcerative Colitis (d) Melanosis (e) Non-adv. polyp (f) Flat polyp

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

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

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

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

2. GASTROINTESTINAL ENDOSCOPY

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

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

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

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

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

3. EIR ARCHITECTURE

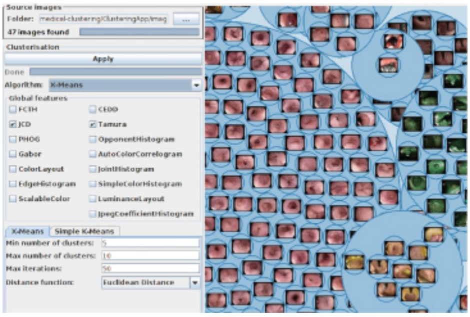
Based on the two target use-cases, the main objectives of the EIR system are (i) easy to use, (ii) easy to extend to different diseases, (iii) real-time handling of multimedia content, (iv) being able to be used as a live system, and (v) high classification performance with minimal false-negative classification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

3.1. Annotation Sub-system

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

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

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



Source images
Folder: medical-clustering/ClusteringApp/imag
47 images found

Clusterisation
Done
Algorithm: X-Means

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

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

Fig. 2. Feature-based clustering annotation.

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

3.2. Detection and Automatic Analysis Sub-system

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

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

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

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

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

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

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

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

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

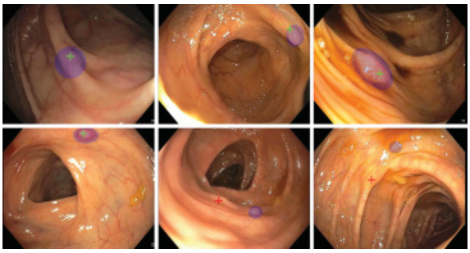


Fig. 3. The localization sub-system marks the possible locations of polyps. The first four show an exact match (ground truth marked with blue ellipses), but the last two are misses.

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

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

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

All the constants and thresholds used in the image rectification and polyp localization algorithms are empirically selected from experimental studies and reflect nuances of the used data. They can be easily adjusted for different datasets, e.g., from another type of endoscope. The image rectification algorithm performs well for all medical imaging artifacts lying outside the main image area. However, it should be improved to be able to detect and remove all pixels that belong to embedded images located anywhere in the frame. This is important for reducing the probability of false positive locations of findings inside of such embedded image regions. The polyp localization algorithm performs well for the used dataset and does not require training data for the detection. An example for the localization output with one possible polyp location is shown in Figure 3.

3.3. Visualization and Computer-Aided Diagnosis Sub-System

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

4. SYSTEM EVALUATION

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

4.1. Detection Accuracy

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2. Localization Accuracy

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

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

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

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

4.3. MICCAI Challenge

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

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

Table V. MICCAI Polyp *localization* Challenge

Participant	True Pos.	False Pos.	False Neg.	Prec.	Recall	F1 score
UNS-UCLAN	48	481	148	9.07	24.49	18.28
CuMedVis	31	167	165	15.75	15.81	15.77
CVC	33	163	163	16.84	16.84	16.84
Our EIR System	46	723	150	5.98	23.47	14.81
RUS	65	1,558	131	4.00	33.16	13.50
SNU	8	188	188	4.08	4.08	4.08

Table VI. MICCAI Polyp *Detection Latency* Challenge

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

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

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

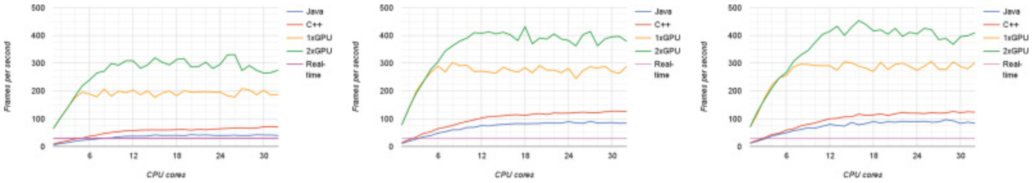
4.4. System Performance

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

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

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

Index	Frames	Total time in seconds	Time per frame in ms
*D*1	3,871	89.78	23.1
*D*2	14,909	178.55	11.9
*D*3	29,818	231.75	7.7
*D*4	100,000	782.351	7.8

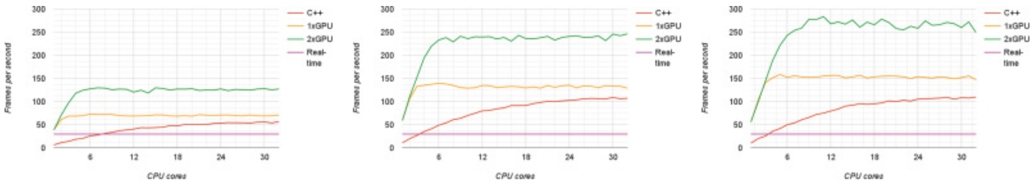


(a) Video resolutions of 1920 × 1080. (b) Video resolutions of 856 × 480. (c) Video resolutions of 712 × 480.

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

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

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



(a) Video resolutions of 1920 × 1080.

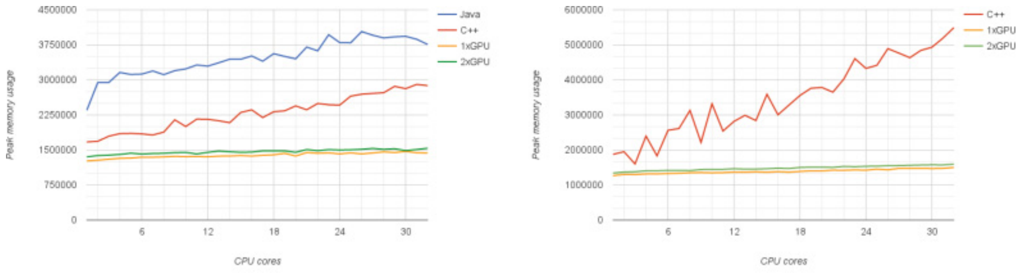
(b) Video resolutions of 856 × 480.

(c) Video resolutions of 712 × 480.

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

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

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



(a) Detection sub-system.

(b) Localisation sub-system.

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

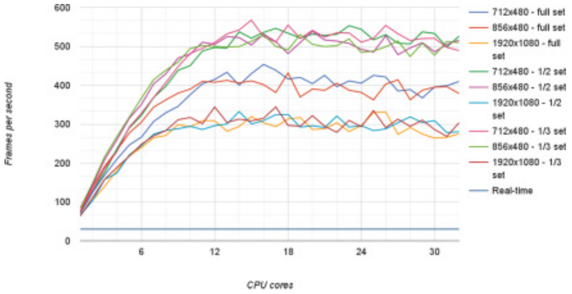


Fig. 7. Performance influence of different training data sizes for 1/2 and 1/3 of the original size.

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

4.4.3. Distributed Processing Experiments. To investigate the performance on distributed hardware for the detection sub-system, some initial experiments on Amazon AWS EC2 instances were conducted. On a *c4.8xlarge* instance (Intel Xeon E5-2666 with 36 virtual CPUs), we were able to classify a video (MPEG-4) with 1,924 frames and a resolution of 1920 × 1080 using the JCD and Tamura features in 29.377 seconds with 65.5 FPS. When classifying data from a raw video file, the processing time increased to 39.599 seconds with 48.6FPS. When reading the data from a Windows media video (wmv) file, the processing time increased to 40.452 seconds with 47.6FPS. The *c4.8xlarge* instance is the most powerful instance offered by Amazon. Therefore, we conducted the same experiments on a less powerful *c4.4xlarge* instance (Intel Xeon E5-2666 with 16 virtual CPUs). Using this instance, we were able to process the MPEG-4 video data in 60.19 seconds with 31.97FPS, the wmv file in 81.17 seconds with 23.7FPS and the raw video file in 79.718 seconds with 24.14FPS. This experiment shows that our system can be distributed, but using the given Amazon hardware, it did not really improve the performance when distributing the workload between several nodes. On the other hand, the performance using only local heterogeneous architectures easily meets the requirements, reducing the need for multi-machine distribution (for now).

5. REAL-WORLD USE-CASES

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

5.1. Live System

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

5.2. Wireless VCE

The multi-sensor VCE is swallowed in order to visualize the GI tract for subsequent detection and diagnosis of GI diseases. Thus, in the future, people may be able to buy VCEs at the pharmacy, and connect and deliver the video stream from the GI tract to the phone over a wireless network. The video footage can be processed in the phone or delivered to our system, which finally analyzes the video automatically. In the best case, the first screening results are available within 8 hours after swallowing the VCE, which is the time the camera typically spends traversing the GI tract. The current VCEs have a low resolution of 256×256 with 3-30FPS (adaptive frame rate with a feedback loop from the receiver to the transmitter). They do not have optimum lighting, making it more challenging to analyze small details in the images. Nevertheless, ongoing work tries to improve the state-of-the-art technology, which will make it possible to use the methods and algorithms developed for colonoscopes also for VCEs [25]. In the case of the colon, accuracy of existing methods is far below the required precision and recall, and the processing of the algorithms does not scale in terms of high-volume data. Each type of disease or irregularity requires interaction between medical researchers dictating what the system must learn to detect, image processing researchers investigating detection or summarization algorithms, hardware developers to develop/produce/research sensors, and distributed processing researchers in order to scale the data analytics of the sensor data. For other scenarios, like in the upper part of the GI tract, there will be similar challenges and corresponding interaction between research disciplines. There are large challenges with respect to accuracy (precision and recall), scale of the processing, and hardware data quality because of different manufacturers (Olympus and Given are the market leaders). The aim is to be a major contributor in the area of medical imaging and sensor processing in the GI tract, as well as storing, processing, and analyzing this type of data.

6. RELATED WORK

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

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

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

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

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

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

†Reported only area under the curve (AUC) instead of sensitivity.

‡Detection and localization performed together. Detection performance alone is around 300FPS and for localization around 100FPS.

and partial systems have been proposed and have achieved, at first glance, promising results in their respective testing environments. However, in some cases, it is unclear how well the approach would perform as a real system used in hospitals. Most of the research conducted in this field uses rather small amounts of training and testing data, making it difficult to generalize the methods beyond the specific dataset and test scenarios. Therefore, overfitting for the specific datasets can be a problem and can lead to unreliable results. Table VIII presents a summary of the most relevant approaches in colonoscopies and polyp detection. The last row of the table shows our approaches' performance to give a comparison. The first approach from Wang et al. [62] is the most recent and best working one in the field of polyp detection. A list of more related work can be found in their paper. As one can see in Table VIII, different methods provide different metrics for measuring the performance and use different datasets for training and testing. Moreover, almost all of them focus on polyp detection. Mamonov et al. [35] presented an algorithm for a binary classifier to detect polyps in the colon. The method is called binary classification with pre-selection, and it aims at reducing the amount of frames that need to be manually inspected. The sensitivity of the algorithm with regards to single input frames is significantly lower and only reaches 47%. A similar approach is presented by Hwang et al. [20]. This approach also focuses on shape, in particular on ellipses, which is a common shape for a polyp. Using this method, a frame is first segmented into regions by a watershed-based image segmentation algorithm. This algorithm is based on the observation that polyps are spherical or hemispherical geometric elevations on the surrounding mucosa. Similar to Mamonov et al. [35], they assume that multiple frames are available for one polyp and that a certain number of false negatives is acceptable in order to balance the number of false positives. The correctness of this assumption depends strongly on the frame rate of the camera that is used for recording the video. As mentioned in the introduction, the best working and complete system in the well-researched polyp detection field is Polyp-Alert [62], which is able to give near real-time feedback during colonoscopies. This approach is also listed as number one in Table VIII. The system can process 10 frames per second and uses visual features and a rule-based classifier to detect the edges of polyps. Further, they distinguish between clear frames and polyp frames in their detection. The researchers report a performance of 97.7% correctly detected polyps based on their dataset, which consists of 52 videos recorded using different colonoscopes. Unfortunately, the dataset is not publicly available, and therefore, an exact detection performance comparison is not possible. Compared to our system, this system seems to reach higher detection accuracy, but it appears that our system is faster in terms of processing time per frame and can therefore detect polyps in real time. A comparison using the same hardware and full-length videos is currently to be carried out together with the developers of Polyp-Alert. Furthermore, our system is not designed and restricted to detect only polyps, and can be expanded to any possible disease if we have the correct training data. Another recent approach not limited to polyps is presented by Nawarathna et al. [36] describing a method to detect bleeding, but also polyps in colonoscopy videos.

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

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

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

7. CONCLUSION

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

REFERENCES

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

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

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

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

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

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

[7] Da-Chuan Cheng, Wen-Chien Ting, Yung-Fu Chen, Qin Pu, and Xiaoyi Jiang. 2008. Colorectal polyps detection using texture features and support vector machine. In *Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry*. Springer, 62–72.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[44] Michael Riegler, Mathias Lux, Carsten Griwodz, Concetto Spampinato, Thomas de Lange, Sigrun L. Eskeland, Konstantin Pogorelov, Wallapak Tavanapong, Peter T. Schmidt, Cathal Gurrin, Dag Johansen, Håvard Johansen, and Pål Halvorsen. 2016. Multimedia and medicine: Teammates for better disease detection and survival. In *Proc. of ACM MM*. 968–977.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Received March 2016; revised March 2017; accepted April 2017