Medico Multimedia Task at MediaEval 2020:
Automatic Polyp Segmentation

Debesh Jha[1,2], Steven A. Hicks[1,3], Krister Emanuelsen[1], Håvard Johansen[2]
Dag Johansen[2], Thomas de Lange[4,5,6], Michael A. Riegler[1], Pål Halvorsen[1,3]
[1] SimulaMet, Norway   [2] UiT The Arctic University of Norway   [3] Oslo Metropolitan University, Norway
[4] Augere Medical AS, Norway   [5] Sahlgrenska University Hospital, Sweden   [6] Bærum Hospital, Norway

ABSTRACT

Colorectal cancer is the third most common cause of cancer worldwide. According to Global cancer statistics 2018, the incidence of colorectal cancer is increasing in both developing and developed countries. Early detection of colon anomalies such as polyps is important for cancer prevention, and automatic polyp segmentation can play a crucial role for this. Regardless of the recent advancement in early detection and treatment options, the estimated polyp miss rate is still around 20%. Support via an automated computer-aided diagnosis system could be one of the potential solutions for the overlooked polyps. Such detection systems can help low-cost design solutions and save doctors time, which they could for example use to perform more patient examinations. In this paper, we introduce the 2020 Medico challenge, provide some information on related work and the dataset, describe the task and evaluation metrics, and discuss the necessity of organizing the Medico challenge.

1 INTRODUCTION

The goal of *Medico automatic polyp segmentation challenge* the benchmarking of polyp segmentation algorithms on new test images for automatic polyp segmentation that can detect and mask out polyps (including irregular, small or flat polyps) with high accuracy. The main goal of the challenge is to benchmark different computer vision and machine learning algorithms on the same dataset that could promote to build novel methods which could be potentially useful in clinical settings. Moreover, we emphasize on robustness and generalization of the methods to solve the limitations related to data availability and method comparison. The detailed challenge description can be found here https://multimediaeval.github.io/editions/2020/tasks/medico/.

After three years of organizing the Medico Multimedia Task [6, 17, 18], we present the fourth iteration in the series. With a focus on assessing human semen quality last year [6], this year we build on the 2017 [18] and 2018 [17] challenges of automatically detecting anomalies in video and image data from the GI tract. We introduce a new task for automatic polyp *segmentation*. In the prior gastrointestinal (GI) challenges, we classified the images into various classes. We are now interested in identifying each pixel of the lesions from the provided polyp images in this challenge.

The task is important because colorectal cancer (CRC) is the third most leading cause of cancer and fourth most prevailing strain in terms of cancer incidence globally [2]. Regular screening through
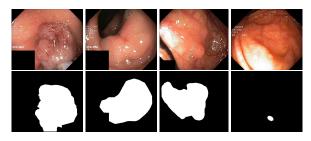
Figure 1: Polyps and corresponding masks from Kvasir-SEG

colonoscopy is a prerequisite for early cancer detection and prevention of CRC. Regardless of the achievement of colonoscopy examinations, the estimated polyp miss rate is still around 20% [12], and there are large inter-observer variabilities [13]. An automated computer-aided diagnosis (CADx) system detecting and highlighting polyps could be of great help to improve the average endoscopist performance.

In recent years, convolutional neural networks (CNNs) have advanced medical image segmentation algorithms. However, it is essential to understand the strengths and weaknesses of the different approaches via performance comparison on a common dataset. There are a large number of available studies on automatic polyp segmentation [3–5, 8, 9, 11, 14, 20]. However, most of the conducted studies are performed on a restricted dataset which makes it difficult for benchmarking, algorithm development and reproducible results. Our challenge is utilizing the publicly available Kvasir-SEG dataset [10]. The entire Kvasir-SEG dataset is used for training and an additional and unseen test dataset for benchmarking the algorithms.

In summary, the Medico 2020 challenge can support building future systems and foster open, comparable and reproducible results where *the objective of the task is to find efficient solutions automatic polyp segmentation*, both in terms of pixel-wise accuracy and processing speed.

For the clinical translation of technologies, it is essential to design methods on multi-centered and multi-modal datasets. We have recently released several gastrointestinal endoscopy [1, 15, 16], wireless capsule endoscopy [19], endoscopic instrument [7], and polyp datasets [10]. Thus, we have put in significant effort to address the challenges related to lack of public available datasets in the field of GI endoscopy.

2 DATASET

The Kvaris-SEG [10] training dataset can be downloaded from https://datasets.simula.no/kvasir-seg/. It contains 1,000 polyp images and
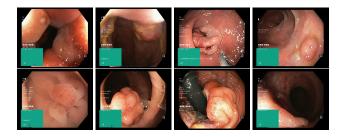
Figure 2: Examples polyps from the test images

their corresponding ground truth mask as shown in Figure 1. The dataset was collected from real routine clinical examinations at Bærum Hospital in Norway by expert gastroenterologists. The resolution of images varies from 332×487 to 1920×1072 pixels. Some of the images contain a green thumbnail in the lower-left corner of the images showing the scope position marking from the ScopeGuide (Olympus) (see Figure 2). We annotate another separate dataset consisting of 160 new polyp images and use the resulting dataset as the test set to benchmark the participants' approaches. Figure 2 shows some examples of test images used in the challenge.

3 TASK DESCRIPTION

The participants are invited to submit their solutions for the two following tasks: segmentation and efficiency (speed).

3.1 The automatic polyp segmentation task

This task invites participants to develop new algorithms for segmentation of polyps. The main focus is to develop an efficient system in terms of diagnostic ability and processing speed and accurately segment the maximum polyp area in a frame from the provided colonoscopic images.

There are several ways to evaluate the segmentation accuracy. The most commonly used metrics by the wider medical imaging community are the correct **Dice similarity coefficient (DSC)** or overlap index, and the **mean Intersection over Union (mIoU)**, also known as the Jaccard index. In clinical applications, the gastroenterologists are interested in pixel-wise detail information extraction from the potential lesions. The metrics such as DSC and mIoU are used to compare the pixel-wise similarity between the predicted segmentation maps and the original ground truth of the lesions.

The DSC is a metric for comparison of the similarities between two given samples. If tp, tn, fp, and fn represent the number of true positive, true negative, false positive and false negative per-pixel predictions for an image, respectively, then the DSC is given as

$$\text{DSC} = \frac{2 \cdot tp}{2 \cdot tp + fp + fn}$$

Furthermore, the IoU is then defined as the ratio of intersection of two metrics over a union of two corresponding metrics. The mean IoU computes IoU of each semantic class of an image and calculate the mean over each classes. The IoU is defined as:

$$\text{IoU} = \frac{tp}{tp + fp + fn}$$

Moreover, in the polyp image segmentation task (i.e., a binary segmentation task), **precision** (positive predictive value) shows over-segmentation, and **recall** (true positive rate) shows under-segmentation. Over-segmentation means that the predicted image covers more area than the ground truth in some part of the frame. The under-segmentation implies that the algorithm has predicted less polyp content in some portion of the image compared to its corresponding ground truth. We also encourage participants to calculate precision and recall, and these are given by:

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

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

The main metric for evaluation and ranking of the teams is **mIoU**. There is a direct correlation between mIoU and DSC. Therefore, we have only used one metric. If the teams have the same mIoU values, then the teams will be further evaluated on the basis of the higher value of the DSC. For the evaluation, we ask the participants to submit the predicted masks in a zip file. The resolution of the predicted masks must be equal to the test images.

3.2 The algorithm speed efficiency task

Real-time polyp detection is required for live patient examinations in the clinic. It can gain gastroenterologist attention to the region of interest. Thus, we also ask participants to participate in the efficiency task. The algorithm efficiency task is similar to the previous task, but it puts a stronger emphasis on the algorithm's speed in terms of frames-per-second.

Submissions for this task will be evaluated based on both the algorithm's speed and segmentation performance. The segmentation performance (the segmentation accuracy) will be measured using the same **mIoU** metric as described above for the first task, whereas speed will be measured by **frames-per-second (FPS)** according to the following formula:

$$FPS = \frac{\#frames}{sec}$$

For this task, we require participants to submit their proposed algorithm as part of a Docker image so that we can evaluate it on our hardware. We evaluate the performance of the algorithm on the Nvidia GeForce GTX 1080 system. For the team ranking, we set a certain mIoU as threshold for considering it as a valid efficient segmentation solution and rank according to the FPS.

4 DISCUSSION AND OUTLOOK

Currently, there is a growing interest in the development of CADx systems that could act as a second observer and digital assistant for the endoscopists. Algorithmic benchmarking is an efficient approach to analyze the results of different methods. A comparison of different approaches can help us to identify challenging cases in the data. We then can discriminate the image frames into simple, moderate, and challenging images. Later on, we can target to develop models on the challenging images that are usually missed out during a routine examination to design better CADx systems. We hope that this approach would help us to design better performing algorithms/models that may increase the efficiency of the health system.

REFERENCES

[1] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data* 7, 1 (2020), 1–14.

[2] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 68, 6 (2018), 394–424.

[3] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. 2020. Pranet: Parallel reverse attention network for polyp segmentation. *arXiv preprint arXiv:2006.11392* (2020).

[4] Yunbo Guo, Jorge Bernal, and Bogdan J Matuszewski. 2020. Polyp Segmentation with Fully Convolutional Deep Neural Networks—Extended Evaluation Study. *Journal of Imaging* 6, 7 (2020), 69.

[5] Yun Bo Guo and Bogdan Matuszewski. 2019. GIANA Polyp Segmentation with Fully Convolutional Dilation Neural Networks. In *Proc. of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. 632–641.

[6] Steven Hicks, Michael Riegler, Pia Smedsrud, Trine B Haugen, Kristin Ranheim Randel, Konstantin Pogorelov, Håkon Kvale Stensland, Duc-Tien Dang-Nguyen, Mathias Lux, Andreas Petlund, et al. 2019. Acm multimedia biomedia 2019 grand challenge overview. In *Proc. of the ACM International Conference on Multimedia*. 2563–2567.

[7] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven Hicks, Vajira Thambawita, Riegler Michael A Garcia-Ceja, Enrique, Lange Thomas de, Peter T. Schmidt, Johansen Håvard, Dag Johansen, and Halvorsen Pål. 2021. Kvasir-Instrument: Diagnostic and Therapeutictool Segmentation Dataset in Gastrointestinal Endoscopy. In *Proc. of International Conference on Multimedia Modeling*.

[8] Debesh Jha, Sharib Ali, Håvard D. Johansen, Dag Johansen, Jens Rittscher, Michael A. Riegler, and Pål Halvorsen. 2020. Real-Time Polyp Detection, Localisation and Segmentation in Colonoscopy Using Deep Learning. *arXiv preprint arXiv:2006.11392* (2020).

[9] Debesh Jha, Michael A Riegler, Dag Johansen, Pål Halvorsen, and Håvard D Johansen. 2020. DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. In *Proc. of International Symposium on Computer-Based Medical Systems*. 558–564.

[10] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-SEG: A segmented polyp dataset. In *Proc. of International Conference on Multimedia Modeling*. 451–462.

[11] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. 2019. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In *Proc. of International Symposium on Multimedia*. 225–230.

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

[13] Nadim Mahmud, Jonah Cohen, Kleovoulos Tsourides, and Tyler M Berzin. 2015. Computer vision and augmented reality in gastrointestinal endoscopy. *Gastroenterology report* 3, 3 (2015), 179–184.

[14] Tanvir Mahmud, Bishmoy Paul, and Shaikh Anowarul Fattah. 2020. PolypSegNet: A Modified Encoder-Decoder Architecture for Automated Polyp Segmentation from Colonoscopy Images. *Computers in Biology and Medicine* (2020), 104119.

[15] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, et al. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proceedings of the ACM on Multimedia Systems Conference*. 170–174.

[16] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. 2017. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proc. of the ACM on Multimedia Systems Conference*. 164–169.

[17] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Steven Hicks, Kristin Ranheim Randel, Duc Tien Dang Nguyen, Mathias Lux, Olga Ostroukhova, and Thomas de Lange. 2018. Medico multimedia task at MediaEval 2018. In *Proc. of MediaEval 2018 CEUR Workshop*.

[18] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Carsten Griwodz, Thomas Lange, Kristin Randel, Sigrun Eskeland, Duc Tien Dang Nguyen, Mathias Lux, and Concetto Spampinato. 2017. Multimedia for medicine: the medico task at Mediaeval 2017. In *Proc. CEUR Worksh. Multim. Bench. Worksh.*

[19] Pia H Smedsrud, Henrik L Gjestang, Oda O Nedrejord, Espen Næss, Vajira Thambawita, Steven Hicks, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. 2020. Kvasir-Capsule, a video capsule endoscopy dataset.

(2020).

[20] Pu Wang, Xiao Xiao, Jeremy R Glissen Brown, Tyler M Berzin, Mengtian Tu, Fei Xiong, Xiao Hu, Peixi Liu, Yan Song, Di Zhang, et al. 2018. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nature biomedical engineering* 2, 10 (2018), 741–748.