Medico Multimedia Task at MediaEval 2018

Konstantin Pogorelov[1,2], Michael Riegler[4,2], Pål Halvorsen[4,2], Steven Alexander Hicks [4]
Kristin Ranheim Randel[2,3], Duc-Tien Dang-Nguyen[5]
Mathias Lux[6], Olga Ostroukhova[7], Thomas de Lange[2]
[1]Simula Research Laboratory, Norway [2]University of Oslo, Norway [3]Cancer Registry of Norway
[4]Simula Metropolitan Center for Digital Engineering, Norway [5]University of Bergen, Norway
[6]University of Klagenfurt, Austria
[7]Research Institute of Multiprocessor Computation Systems n.a. A.V. Kalyaev, Russia
konstantin@simula.no,michael@simula.no

ABSTRACT

The *Medico: Multimedia for Medicine Task*, running for the second time as part of MediaEval 2018, focuses on detecting abnormalities, diseases, anatomical landmarks and other findings in images captured by medical devices in the gastrointestinal tract. The task is described, including the use case and its challenges, the dataset with ground truth, the required participant runs and the evaluation metrics.

1 INTRODUCTION

The 2018 version of the Medical Multimedia task (Medico) continues tackling the challenge of utilizing multimedia data collected in hospitals. As a first goal, the task focuses on efficient (in terms of resources, time and accuracy) automatic detection of anatomical findings using images and videos. In addition, automatic report creation / content summarization of videos is provided as a possible sub-task.

In comparison to other medical imaging related challenges, e.g., [1, 9], the task differs in that the final goal is the inclusion of medical experts as users, and no use of additional information (multimodal analysis). The main differences are that Medico (i) provides only multimedia data (videos and images) and no medical imaging data (CT scans, etc.), (ii) asks for using as little training data as possible, (iii) evaluates the approaches also regarding processing time, and (iv) asks to perform automatic report generation and summarization of patients assessments.

In the medial field, most of the contributions nowadays are made by the communities focusing on visual analysis such as computer vision and medical imaging researchers. These are important contributors to the overall challenge to improve the medical services, but still only a small part of the big picture. The medical field holds a lot of different multimedia related research problems that are not yet tackled. It is also important to see the medical field not just as another use case for already existing multimedia systems and algorithms, but rather as a field that needs its own research. An algorithm designed for detecting cars or to summarize personal image collections will not be able to tackle the domain-specific challenges of medical multimedia data [8]. To be able to significantly improve the healthcare system, it is necessary to go beyond image and video based analysis, make serious use of the multitude of additional information sources including sensors and temporal information and to finally create multimedia systems and applications that provide fast (real-time) feedback [7] and include the patients and doctors as users [3, 10, 11].

In this respect, the Medico task is designed to encourage multimedia researchers to apply their skills and knowledge in the field of medicine with a low starting threshold by providing open medical-related data and expert knowledge. The task was created in collaboration with medical experts to ensure usefulness and real world reference. Among many different medical sub-fields, endoscopic examinations of the gastrointestinal (GI) tract have been chosen. Early detection of abnormalities and diseases in the GI tract can significantly improve the chance of successful treatment and survival of patients and also reduce medical costs. This is particularly the case for colorectal cancer (in the large bowel) or its cancer precursors (colorectal polyps), which can be detected through colonoscopy procedures [12]. The challenge is, however, that both medical experts and machines currently fail to achieve sufficient detection rates [8]. There is a need for image and video processing, analysis, information search and retrieval, in combination with other sensor data and assistance from medical experts, and it all needs integration [4].

The Medico task challenges researchers beyond computer vision and medical imaging to show the potential of multimedia research outside well-known scenarios like analysis of content on YouTube, Twitter or Facebook. The task provides, for the medical field, a large publicly available dataset containing videos and images from the GI tract showing different diseases, anatomical landmarks and other findings including normal GI tract tissue. The dataset provided this year consists of newly collected and annotated images as well as images from the previous published datasets called Kvasir [6] and Nerthus [5]. The ground truth is provided by medical experts (specialists in GI endoscopy) annotating the dataset, and the data is split into training and test data. Based on the provided dataset, the participants are asked to solve three subtasks where the two first are mandatory, and the one last is optional:

(i) classify diseases with as few images in the training dataset as possible;
(ii) solve the classification problem in a fast and efficient way on "normal" PC hardware;
(iii) generate a text-report for a medical doctor that can fulfill existing demands for documentation of endoscopic procedures.

Tackling the task can be addressed by leveraging techniques from multiple multimedia-related disciplines, including (but not limited

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

to) machine learning (classification), multimedia content analysis and multimodal fusion.

2 DATASET DETAILS

The dataset provided consists of 14,033 GI tract images with different resolutions (from 720x576 up to 1920x1072 pixels) that are annotated and verified by medical doctors (experienced endoscopists) for the ground truth. It includes 16 classes, 8 more than the previous years version of the task, showing anatomical landmarks, pathological and normal findings or endoscopic procedures in the GI tract, with different numbers of images for each class, split into development (training) and testing sets. The anatomical landmarks are *normal-z-line*, *normal-pylorus*, *normal-cecum*, *retroflex-rectum*, *retroflex-stomach*, while the pathological findings include *esophagitis*, *polyps* and *ulcerative-colitis*. The pre-, while- and post-surgery findings are the *dyed-lifted-polyps*, the *dyed-resection-margins* and the *instruments*. Additional classes include normal tissue with or without stool contamination are the *colon-clear*, the *stool-inclusions* and the *stool-plenty*, as well as some image classes are not usable for diagnosis, namely the *blurry-nothing* and the *out-of-patient*.

2.1 Dataset content

The development and the test datasets consist of 5,293 images and 8,740 images, respectively, stored in two archives each: images archive and features archive. Both datasets are heavily unbalanced in terms of number of samples per class which reflects the real practice in hospitals when doctors tend to collect only selected classes of images giving no attention to, for example, normal findings and routine objects like stool. Thus, the number of images per class in the sets can vary from few up to thousands images.

In the development dataset, the images are stored in the separate folders named according to the name of the classes that images belong to. In the test dataset, all the images are stored in one folder. The image files are encoded using JPEG compression. The encoding settings can vary across the dataset. Furthermore, the features archive contains pre-extracted visual feature descriptors for all the images of the dataset that can optionally be used by the participants. The extracted visual features are JCD, Tamura, ColorLayout, EdgeHistogram, AutoColorCorrelogram and PHOG [2]. The extracted visual features are stored in the separate folders and text files with ".features" extension named according to the name and the path of the corresponding image files. Each file consists of eight lines, one line per feature, and a line consists of a feature name separated from the feature vector by colon. Each feature vector consists of a corresponding number of floating point values separated by commas.

3 AUTOMATIC REPORT GENERATION

For the automatic report generation, we use three videos depicting diseases or findings that can be found in the development dataset. The goal is to generate reports summarizing the three videos for medical experts having an automatic report generation in mind. The level of complexity of the summary or report is up to the participant and will be assessed by a jury of three medical experts in terms of usefulness for a real use in hospitals.

4 EVALUATION METRICS AND TASKS

For the evaluation of detection accuracy, we use several standard metrics (more detailed descriptions on the task web-page). The officially reported metric for evaluating the multi-class classification is the Matthews correlation coefficient (MCC). In case of equal MCC values, we use *the amount of training data* that has been used to achieve good results and the *speed* (processing performance) of the classification. We also evaluate the *the amount of training data* and the *speed* separately in order to verify classification algorithms' suitability to be used in real-world conditions.

For the evaluation, the participants must submit one run for the required subtask defined below. Additionally, they optionally can submit three more for any of the described subtasks, i.e., participants can submit up to five runs in total.

Required subtask 1: The *detection subtask* is a task for multi-class classification of diseases in the GI tract. Participants have to use visual information in the provided dataset where the goal is to maximize the algorithm's performance in terms of detection accuracy, where amount of training data is also taken into account. Detection is evaluated and ranked using MCC. The amount of used training data and processing speed can also be taken into account as described earlier.

Optional subtask 1: The *efficient detection subtask* addresses the speed of the classification on a "normal" PC hardware. The classification of diseases has to be achieved as fast as possible in terms of data processing using any computation speed-up techniques. The goal is to find a balance between the algorithm's performance in terms of detection accuracy and the performance in terms of data processing speed, i.e., keeping in mind that the problem area requires fast processing for real-time feedback and diagnosis. For the evaluation and ranking, the processing time is weighted by the MCC.

Optional subtask 2: The experimental *report generation subtask* asks the participants to automatically create a report or summary for a medical doctor describing the detection results for three video cases. The report is assessed and ranked manually by three of our medical partners in terms of usefulness in the medical context in hospitals.

5 DISCUSSION AND OUTLOOK

The task itself can be seen as very untypical and challenging. Due to the importance of the use case, we hope to motivate researchers usually not working in the medical field to present their approaches. Performing research that can have societal impact will be an important part of multimedia research in the future. We hope that the Medico task can help to raise awareness of the topic, but also provide an interesting and meaningful use case to researchers interested in this direction.

REFERENCES

[1] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba Garcia Seco de Herrera, Cathal Gurrin, Bayzidul Islam, Vassili Kovalev, Vitali Liauchuk, Josiane Mothe, Luca Piras, Michael Riegler, and Immanuel Schwall. 2017. Overview of ImageCLEF 2017: Information extraction from images. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference*

of the CLEF Association, CLEF 2017 (LNCS 10439). Springer.

[2] Mathias Lux and Savvas A Chatzichristofis. 2008. Lire: lucene image retrieval: an extensible java cbir library. In *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 1085–1088.

[3] Konstantin Pogorelov, Sigrun Losada Eskeland, Thomas de Lange, Carsten Griwodz, Kristin Ranheim Randel, Håkon Kvale Stensland, Duc-Tien Dang-Nguyen, Concetto Spampinato, Dag Johansen, Michael Riegler, and others. 2017. A holistic multimedia system for gastrointestinal tract disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSYS)*. ACM, 112–123.

[4] Konstantin Pogorelov, Olga Ostroukhova, Mattis Jeppsson, Håvard Espeland, Carsten Griwodz, Thomas de Lange, Dag Johansen, Michael Riegler, and Pål Halvorsen. 2018. Deep Learning and Hand-crafted Feature Based Approaches for Polyp Detection in Medical Videos. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE.

[5] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 170–174.

[6] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSYS)*. ACM, 164–169.

[7] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Peter Thelin Schmidt, Carsten Griwodz, Dag Johansen, Sigrun Losada Eskeland, and Thomas de Lange. 2016. GPU-accelerated real-time gastrointestinal diseases detection. In *Proceeding of the IEEE International Symposium onComputer-Based Medical Systems (CBMS)*. IEEE, 185–190.

[8] Michael Riegler, Mathias Lux, Carsten Gridwodz, 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 *Proceedings of the 2016 ACM Multimedia Conference (ACM MM)*. ACM, 968–977.

[9] Mauricio Villegas, Henning Müller, Alba García Seco de Herrera, Roger Schaer, Stefano Bromuri, Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Arnau Ramisa, and others. 2016. General overview of imageCLEF at the CLEF 2016 labs. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages (LNCS 9822)*. Springer, 267–285.

[10] Yi Wang, Wallapak Tavanapong, Johnny Wong, JungHwan Oh, and Piet C De Groen. 2011. Computer-aided detection of retroflexion in colonoscopy. In *Proceeding of the 24th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 1–6.

[11] 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.

[12] Sidney Winawer, Robert Fletcher, Douglas Rex, John Bond, Randall Burt, Joseph Ferrucci, Theodore Ganiats, Theodore Levin, Steven Woolf, David Johnson, and others. 2003. Colorectal cancer screening and surveillance: clinical guidelines and rationaleâĂŤupdate based on new evidence. *Gastroenterology* 124, 2 (2003), 544–560.