Multimedia for Medicine: The Medico Task at MediaEval 2017

Michael Riegler[1], Konstantin Pogorelov[1,2], Pål Halvorsen[1,2],
Kristin Ranheim Randel[2,3], Sigrun Losada Eskeland[4], Duc-Tien Dang-Nguyen[5],
Mathias Lux[6], Carsten Griwodz[1,2], Concetto Spampinato[7], Thomas de Lange[3]
[1]Simula Research Laboratory, Norway [2]University of Oslo, Norway [3]Cancer Registry of Norway, Norway
[4]Vestre Viken Hospital Trust, Norway [5]Dublin City University, Ireland [6]University of Klagenfurt, Austria
[7]University of Catania, Italy
michael@simula.no,konstantin@simula.no

ABSTRACT

The *Multimedia for Medicine Medico Task*, running for the first time as part of MediaEval 2017, focuses on detecting abnormalities, diseases and anatomical landmarks in images captured by medical devices in the gastrointestinal tract. The task characteristics are described, including the use case and its challenges, the dataset with ground truth, the required participant runs and the evaluation metrics.

1 INTRODUCTION

The Medico task tackles the challenge of predicting diseases based on multimedia data collected in hospitals with the additional requirements to use as little training data as possible, perform the analysis efficient regarding processing time and to generate automatic text reports (summaries) of the findings. The task differs from well know medical imaging tasks like the ImageClef medical tasks (http://www.imageclef.org/) [1, 7] in the points that it (i) has only multimedia data (videos and images) and no medical imaging data (CT scans, etc.), (ii) asks for using as little training data as possible and (iii) evaluates the approaches also regarding processing time. Furthermore, the automatic generated reports are a novel part of the task, but since it is very hard to evaluate them this subtask is experimental this year.

It is a typical assumption that visual analysis as it is already provided by the computer vision and medical image processing communities today is sufficient to solve healthcare multimedia challenges [6]. Although we concede that these methods are indeed essential contributors to promising approaches, we have come to the understanding that analysing images and videos alone does not solve the challenges in medical fields such as endoscopy or ultrasound, because of the task complexity and the needs of both medical experts and patients. Neither does it make serious use of the multitude of additional information sources including sensors and temporal information [3, 8, 9].

The Medico task is designed to help to improve the health care system through application of multimedia research knowledge and methods to reach the next level of computer and multimedia-assisted diagnosis, detection and interpretation of abnormalities. This is useful in multiple scenarios. For example, in some areas of the human body, such as the gastrointestinal (GI) tract, the detection of abnormalities and diseases in early stages can significantly improve the chance of successful treatment and survival. Through endoscopic examinations (insertion of a camera in the gastrointestinal tract), diseases can be detected visually, even before they become symptomatic. This is particularly the case for colorectal cancer (in the large bowel) or its cancer precursors (colorectal polyps), which can be detected through colonoscopy or capsule endoscopy. The challenge is, however, that both medical experts and machines currently fail to detect all polyps [6]. Moreover, in previous research in this area, computer vision and medical imaging have created visual augmentations of the interior of a body. To automatically detect and locate abnormalities, visual representations are not sufficient. There is a need for image and video processing, analysis, information search and retrieval, in combination with other sensor data and assistance from medical experts, and it all needs integration [5].

Here, participants are asked to look beyond computer vision and medical imaging to show the potential of multimedia research going far beyond well known scenarios like analysis of content on YouTube and Flickr. For this detection task, we provide Kvasir, a large public dataset [4] containing videos and images from the GI tract showing different diseases and anatomical landmarks. 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 this, the participants are asked to solve four subtasks, i.e., the two first are mandatory, and the two last are 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; (iii) run the second task on the same hardware (supported platforms are Linux, macOS and Windows); and (iv) automatically create a text-report for a medical doctor for three video cases. Tackling the task can be addressed by leveraging techniques from multiple multimedia-related disciplines, including (but not limited to) machine learning (classification), multimedia content analysis and multimodal fusion.

2 DATASET DETAILS

The Kvasir dataset[1] [4] consists of 8,000 GI tract images that are annotated and verified by medical doctors (experienced endoscopists) for the ground truth. It includes 8 classes showing anatomical landmarks, pathological findings or endoscopic procedures in the GI tract, i.e., 1000 images for each class, split into 500 for training and 500 for testing. The anatomical landmarks are *Z-line*, *pylorus* and *cecum*, while the pathological findings include *esophagitis*, *polyps* and *ulcerative colitis*. In addition, we provide two set of images related to removal of polyps, the *dyed and lifted polyp* and the *dyed resection margins*. The dataset consists of images with different

[1]http://datasets.simula.no/kvasir/

MediaEval'17, 13-15 September 2017, Dublin, Ireland

resolutions from 720x576 up to 1920x1072 pixels and is organized by sorting them into separate folders named according to the content. Some of the included images have a green sub-picture in the image illustrating the position and configuration of the endoscope inside the bowel, delivered from an electromagnetic imaging system (ScopeGuide, Olympus Europe). This sub-picture may support the interpretation of the image. As mentioned before, the whole dataset is split into two equally sized development and test datasets. Both the development and the test datasets consist of 4,000 images, 500 images for each class stored in two archives: images archive and features archive. 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 stored in one folder. The image files are encoded using JPEG compression. The encoding settings can vary across the dataset, and they reflect the a priori unknown endoscopic equipment settings. Furthermore, the features archive contains the extracted visual feature descriptors for all the images in the images archive. The extracted visual features are the global image features, i.e., JCD, Tamura, ColorLayout, EdgeHistogram, AutoColorCorrelogram and PHOG. Each feature vector consists of a number of floating point values. The size of the vector depends on the feature. The sizes of the feature vectors are: 168 (JCD), 18 (Tamura), 33 (ColorLayout), 80 (EdgeHistogram), 256 (AutoColorCorrelogram) and 630 (PHOG) [2]. The extracted visual features are stored in the separate folders and text files named according to the name and the path of the corresponding image files. Each file consists of six 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. The extension of the extracted visual feature files is ".features".

For the automatic report generation, we use three videos depicting diseases or findings that can be found in the Kvasir dataset. The goal is to generate reports describing the three videos for medical experts having an automatic report generation in mind.

3 EVALUATION METRICS AND TASKS

For the evaluation of detection accuracy, we use several standard metrics (more detailed descriptions on the task web-page). *True positive* represents the number of correctly identified samples. *True negative* shows the number of correctly identified negative samples. *False positive* is the number of wrongly identified samples. *False negative* denotes the number of wrongly identified negative samples. *Recall* (frequently called sensitivity) is the ratio of samples that are correctly identified as positive among all existing positive samples. *Precision* shows the ratio of samples that are correctly identified as positive among the returned samples. *Specificity* represents the ratio of negatives that are correctly identified as negatives. *Accuracy* is the percentage of correctly identified true and false samples. *Matthews correlation coefficient* (MCC) takes into account true and false positives and negatives, and is a balanced measure even if the classes are of very different sizes. *F1 score* is a measure of a test's accuracy by calculating the harmonic mean of the precision and recall. We also evaluate *the amount of training data* that has been used to achieve good results and the *speed* (processing performance) of the classification. For the evaluation, the participants must submit one run for each of the required subtasks 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 subtasks. 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 based on the metrics above (all should be reported), but a ranking is made using MCC and the amount of used training data. The official metric is a multi-class generalization of the MCC. This generalization is called the R_K statistic (for K different classes) and defined in terms of a $K \times K$ confusion matrix. The R_K statistic is in essence a correlation coefficient between the observed and predicted binary classifications for (for K different classes); it returns a value between -1 and $+1$. A coefficient of $+1$ represents a perfect prediction, 0 corresponds to no better than random prediction and a value < 0 indicates disagreement between prediction and observation (the lower negative value corresponds to the stronger disagreement). The minimum negative value of the R_K statistic is between -1 and 0 depending on the true distribution. The maximum value is always $+1$.

The *efficient detection subtask* addresses the speed of the classification. 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, while keeping in mind that the problem area requires real-time processing while lacking data. For the evaluation, the processing time weighted by the detection accuracy.

Optional subtasks. The *efficient detection on same hardware subtask* is the same as the efficient detection subtask above, but all submitted solutions are tested on the same hardware. The organizers run the code provided by the participants on the same hardware, and the evaluation is again based on the processing time weighted by detection accuracy is used.

The experimental *report generation subtask* asks the participants to automatically create a text-report for a medical doctor describing the detection results for three video cases. A definition of what a text report is, what it should contain (list of requirements) and a description of what the medical experts do with the report is provided. The assessment then follows the list of requirements, and the report is assessed manually from two of our medical partners in terms of usefulness in the medical context and if it satisfies existing demands for documentation of endoscopic procedures.

4 DISCUSSION AND OUTLOOK

The task itself can be seen as very challenging, hard to solve and hard to evaluate. Due to its novel use case, we hope to motivate a lot of researchers to have a look into the field of medical multimedia. 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.

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, 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.

[5] 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.

[6] 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.

[7] 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 *Procedings of the International Conference of the Cross-Language Evaluation Forum for European Languages (LNCS 9822)*. Springer, 267–285.

[8] 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.

[9] 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.