Validating convolution neural networks for automatic polyp detection in robotic colonoscopy

P. Brandao[1], E. Mazomenos[1], G. Ciuti[2], F. Bianchi[2], Menciassi[2], P. Dario[2], A. Koulaouzidis[3], D. Stoyanov[1]

[1]Centre for Medical Image Computing, University College London
[2] The BioRobotics Institute, Scuola Superiore Sant'Anna
[3] Endoscopy Unit, The Royal Infirmary of Edinburgh

INTRODUCTION

Colorectal cancer (CRC) is the most frequent malignancy of the gastrointestinal tract and accounts for nearly 10% of all forms of cancer [1]. The survival rate of CRC patients is lower than 7% when the disease reaches an advanced stage, however, in cases of early diagnosis, with successful treatment, it increases to almost 90% [1]. Even though conventional colonoscopy is considered the most effective method for CRC screening and diagnosis, invasiveness, patient discomfort and fear of pain can lead to patient reluctance towards the examination [1,2]. Robotic endoscopic capsules and endoscopes can potentially overcome the drawbacks of pain and discomfort while still facilitating full movement control to the clinician on order to observe the entire endoluminal colonic wall [1].

In any kind of colonoscopy, the success of the exam highly depends on the operator's skills, as a high level of hand-eye coordination is required to examine the majority of the colon wall [1]. In 2012, Leufkens *et al.* reported that missed polyp detection rates in colonoscopy screenings could reach values as high as 25% [3]. One effective way to increase this detection rate is the incorporation of computer-aided diagnostic systems [1].

In this study, we propose an automatic polyp detection system using a deep learning framework that can aid clinicians to detect polyps more accurately in colorectal exams. This approach can be effective in helping endoscopists to detect high risk regions and can additionally be used to reduce observation times for capsule endoscopy (CE) procedures. To our knowledge, this is the first work to apply convolution neural networks (CNNs) for polyp detection in CE images.

MATERIALS AND METHODS

As a preliminary study, we evaluated the ability of the method to detect polyp-like structures in images of a colon phantom obtained from a robotically controlled endoscopic capsule. In this section we describe the experimental setup, the data used and the learning strategies adopted.

Experimental Setup

Our experimental setup comprised a PillCam™ SB3 capsule (Given Imaging, Yokneam, IL), shown in Figure. 1 a), attached to a plastic rigid rod that was controlled by an industrial anthropomorphic robotic arm (Mitsubishi RV-3SB robot, Tokyo, JP - Figure. 1 b)).

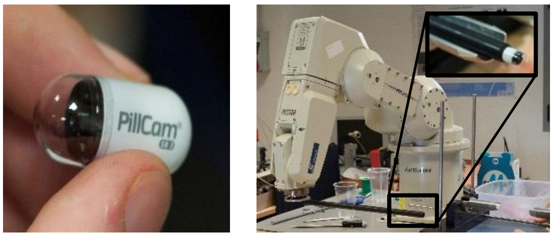

Figure. 1. (a) PillCam™ SB 3 model; (b) Endoscopic capsule and plastic rigid rod attached to an industrial anthropomorphic robotic arm.

Figure. 2 a) shows the deformable and impermeable colon phantom used in this study (Lifelike Biotissue Inc, Ontario, CAN). During tests small coloured pins were inserted along the phantom track to emulate polyps, as shown in Figure. 2 b).

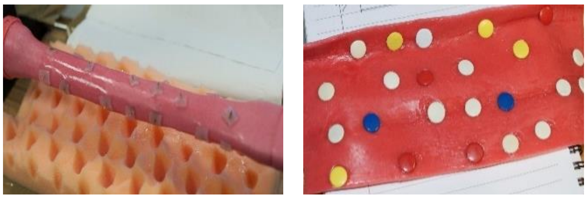

Figure. 2. a) External view of the colon phantom; b) View of the opened phantom with the circular pins attached.

The study was conducted by inserting the capsule, with its camera end facing forward, into the phantom. To emulate a real colon exam, the robotic arm moved the camera horizontally throughout the track of the phantom and back. This procedure was repeated multiple times and all data was wirelessly recorded using Given Imaging software.

Dataset

From the video, containing several forward and backward sweeps of the phantom, 100 frames were randomly selected and extracted. Every visible pin was manually segmented and the data was randomly divided into same size training and testing datasets.

Fully Convolution Neural Network

Motivated by the success reported in general recognition tasks [4,5], we decided to apply CNNs for polyp detection. Even though CNNs are usually formulated to solve classification problems, some approaches were able to use CNNs for dense prediction by labelling each

**The research leading to these results has received funding from the European Community's Horizon 2020 Programme (H2020/2014-2020) under Grant Agreement num. 688592 (EndoVESPA Project).*

pixel with the class of its enclosing object. This is usually achieved by post-processing super-pixel projection, multi-scale approaches or patch-wise training. In this study, we use the 8-strided variant of the VGG net proposed by Long *et al.*, where dense prediction is obtained by in-network deconvolution layers that are also learned during training [6][6]. This results in a very fast and effective upsampling that achieved state-of-the-art in semantic segmentation of the PASCAL VOC dataset [5]. Training a deep network from raw data is not practical considering the time and computational power required. In addition, it may not capture the full image complexity of the scene. Because of this, we fine-tune pre-trained layers on the ImageNet dataset [4] by backpropagation using our training data. The last fully connected layer of the CNN was replaced with a binary classification layer. Training was done iteratively by SGD with momentum using a batch size of 50 images and fixed learning rates of 10^{-11}. Due to memory constrains, all images were resized to 180 × 180 pixels. All processing was done using Caffe [7] in a NVIDIA Quadro K4000 GPU.

RESULTS

We evaluate the prediction results in three separate points during training. The resultant binary segmentations of each network were used to calculate the performance metrics on the testing dataset. Precision and recall results are presented in Table 1.

Table 1. Average pixel-wise precision and recall of the networks in the testing dataset during different training stages. fCNN-800, fCNN-900 and fCNN-1000 represents the network after 800, 900 and 1000 training iterations, respectively

Network	Precision	Recall
fCNN-800	0.55	0.95
fCNN-900	0.77	0.84
fCNN-1000	0.59	0.95

Figure. 3, illustrates the segmentation output of the networks for the same image.

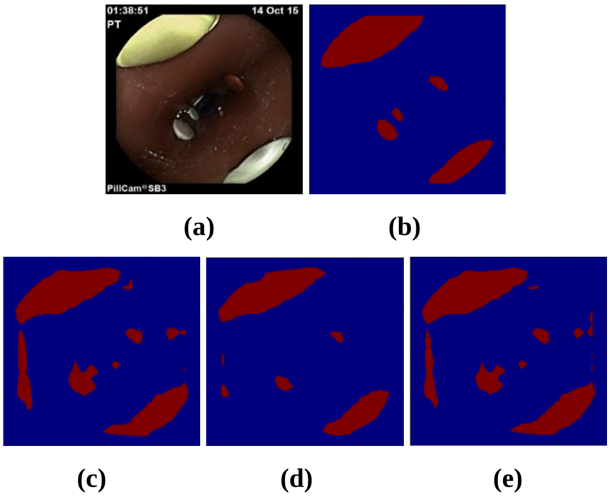

Figure. 3. (a) Example of a testing image; (b) Manual segmentation of the pins; (c), (d) and (e) binary segmentations produce by fCNN-800, fCNN-900 and fCNN-1000, respectively.

CONCLUSION AND DISCUSSION

The best polyp segmentation results were obtained by fCNN-900, with a precision and recall of 0.77 and 0.84, respectively. In the earlier training iteration, fCNN-800 has not converged to the best possible solution yet and the prediction results are less accurate. As we continue training, the fCNN-1000 starts to overfit to the training data, and inference ability is lost. The same principle is supported by Figure. 3, where the segmentation of fCNN-900 outperforms the others.

Figure. 3 d) shows that the method slightly underperforms in segmenting smaller pins compared to larger ones. This is expected, as the dense prediction is obtained from the deconvolution upsampling layer, where the minimum resolution is limited by the adopted stride. This effect could be minimized by using a GPU with more memory, which would not impose downscaling of the input images. Furthermore, the considerable amount of false positives suggests that the model struggles to classify small variations in the background. These can easily be corrected with the use of more training data.

In conclusion, we were able to successfully employ CNNs to automatically segment polyp-like structures in images from a robotically controlled endoscopic capsule. Even with a limited amount of training data and GPU memory, the method performs well, which is promising for translation into the clinical setting. Future work involves verifying our results to colonoscopy data using the large datasets available from the 2015 MICCAI sub-challenge on automatic polyp detection [8] and to data from real CE examinations.

REFERENCES

[1] G. Ciuti, R. Caliò, D. Camboni, et al., “Frontiers of robotic endoscopic capsules: a review,” *J. Micro-Bio Robot.*, pp. 1–18, May 2016.

[2] W. C. Leung, D. C. Foo, T. Chan, et al., “Alternatives to colonoscopy for population-wide colorectal cancer screening,” *Hong Kong Med. J.*, Jan. 2016.

[3] A. M. Leufkens, M. G. H. Van Oijen, F. P. Vleggaar, and P. D. Siersema, “Factors influencing the miss rate of polyps in a back-to-back colonoscopy study,” *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.

[4] O. Russakovsky, J. Deng, H. Su, , et al., “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[5] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal Visual Object Classes Challenge: A Retrospective,” *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2014.

[6] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *CVPR*, 2015, vol. 07–12-June, pp. 3431–3440.

[7] Y. Jia, E. Shelhamer, J. Donahue, , et al. “Caffe: Convolutional Architecture for Fast Feature Embedding,” *Proc. ACM Int. Conf. Multimed.*, 2014.

[8] A. Bernal, J., Tajkbaksh, N., Sánchez, F.J., Liang, J., Chen H., Yu, L., Angermann, Q., Romain, et al., “Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge,” *IEEE Trans. Med. Imaging*, 2016.