Towards Automated Colonoscopy Diagnosis: Binary Polyp Size Estimation via Unsupervised Depth Learning

Hayato Itoh[1(✉)], Holger R. Roth[1,2], Le Lu[2], Masahiro Oda[1], Masashi Misawa[3], Yuichi Mori[3], Shin-ei Kudo[3], and Kensaku Mori[1,4,5]

[1] Graduate School of Informatics, Nagoya University, Nagoya, Japan
hitoh@mori.m.is.nagoya-u.ac.jp
[2] AI-Infra, NVIDIA Corporation, Santa Clara, USA
[3] Digestive Disease Center, Showa University Northern Yokohama Hospital, Yokohama, Japan
[4] Information Technology Center, Nagoya University, Nagoya, Japan
[5] Research Center for Medical Bigdata, National Institute of Informatics, Tokyo, Japan

Abstract. In colon cancer screening, polyp size estimation using only colonoscopy images or videos is difficult even for expert physicians although the size information of polyps is important for diagnosis. Towards the fully automated computer-aided diagnosis (CAD) pipeline, a robust and precise polyp size estimation method is highly desired. However, the size estimation problem of a three-dimensional object from a two-dimensional image is ill-posed due to the lack of three-dimensional spatial information. To circumvent this challenge, we formulate a relaxed form of size estimation as a binary classification problem and solve it by a new deep neural network architecture: BseNet. This relaxed form of size estimation is defined as a two-category classification: under and over a certain polyp dimension criterion that would provoke different clinical treatments (resecting the polyp or not). BseNet estimates the depth map image from an input colonoscopic RGB image using unsupervised deep learning, and integrates RGB with the computed depth information to produce a four-channel RGB-D imagery data, that is subsequently encoded by BseNet to extract deep RGB-D image features and facilitate the size classification into two categories: under and over 10 mm polyps. For the evaluation of BseNet, a large dataset of colonoscopic videos of totally over 16 h is constructed. We evaluate the accuracies of both binary polyp size estimation and polyp detection performance since detection is a prerequisite step of a fully automated CAD system. The experimental results show that our proposed BseNet achieves 79.2 % accuracy for binary polyp-size classification. We also combine the image feature extraction by BseNet and classification of short video clips using a long short-term memory (LSTM) network. Polyp detection (if the video clip contains a polyp or not) shows 88.8 % sensitivity when employing the spatio-temporal image feature extraction and classification.

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 611–619, 2018.
https://doi.org/10.1007/978-3-030-00934-2_68

Keywords: Size estimation · Depth estimation · Deep neural networks · Long short-term memory (LSTM) · Polyp detection

1 Introduction

Size information of detected polyps is an essential factor for diagnosis in colon cancer screening. For example, the U.S. guideline for colonoscopy surveillance defines what treatments should be acted after a physicians find polyps with respect to their size estimations [1]. Whether the size of a polyp is over or under 10 mm is important. The guideline [1] defines that patients with only 1 or 2 small (≤10 mm) tubular adenomas with only low-grade dysplasia should have their next follow-up in 5−10 years; and patients with 3 to 10 adenomas, or any adenoma >10 mm, or any adenoma with villous features or high-grade dysplasia will have follow-ups in 3 years. However, polyp size estimation using only colonoscopy is quite difficult even for expert physicians so that automated size estimation techniques would be desirable.

In general, size estimation of a 3D object from a 2D image is an ill-posed problem due to the lack of three-dimensional spatial information. Figure 1 demonstrates the challenge of the polyp-size estimation from a colonoscopic image. Polyps with different diameters from 2 mm to 16 mm may have the similar image sizes or ranges. The image size of a polyp depends on both the true 3D polyp size and the physical distance from colonoscope to the polyp. Our key question is that will the recovered image depth information augmented with original colonoscopic RGB images be helpful for polyp size estimation and detection. Depth maps from monocular colonoscopic RGB images can be computed through unsupervised deep neural network [2,3].

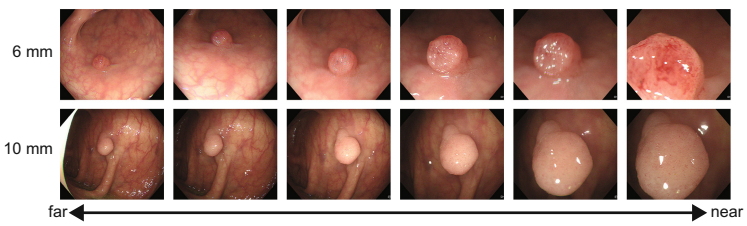


Fig. 1. Examples of polyps on colonoscopic images. Top and bottom rows show images that capture polyps with diameters of 6 mm and 10 mm, respectively. From left to right, columns show images with different (long to short) distances from colonoscope to polyps.

Previous techniques have been proposed for 3D scene reconstruction and camera pose recovery out of 2D images [4,5]. Reference [4] extracts invariant geometry features of rigid objects and complies them to the geometrical constraints of cameras to reconstruct 3D information. In colonoscopy, there is only

one light source where shading based 3D object shape reconstruction [5] is possible. However these 3D reconstruction methods may not work well in colonoscopy due to the non-rigidness and complex textures the colon wall.

We propose a new method for binary polyp size estimation or classification from a single colonoscopic image. The problem of size estimation is relaxed into a binary size classification task according to guidelines [1]. We propose the binary-size estimation network (BseNet) to solve two-category polyp classification. First, BseNet estimates depth maps from three-channel (RGB) colonoscopic images via unsupervised depth recovery convolutional neural networks [2,3], and integrates all channels into RGB-D imagery. Second, RGB-D image features from the newly integrated imagery are extracted. Third, the two-category classification for binary size estimation is performed by classifying these RGB-D image features. Finally, For a complete and automated computer-aided polyp diagnosis system, we exploit the polyp detection performance based on spatio-temporal deep features by leveraging a large dataset of colonoscopic videos.

2 Methods

2.1 Spatio-temporal Video Based Polyp Detection

Before estimating binary polyp sizes, polyp detection is a prerequisite processing step with no de facto standard methods [6,7]. In this paper, we adopt scene classification representation to classify the existence status of polyps in any colonoscopic video sub-clips: as *positive* when at least one polyp exists, or *negative* when there is no polyp. Polyp detection in colonoscope imagery requires the extraction of spatio-temporal image feature from videos. Successive colonoscopic image frames usually include similar objects of the same scene category. In particular, for the positive category, a polyp should appear in successive frames. Therefore, polyp detection as scene classification needs to handle the temporal context in addition to the spatial structure of 2D images.

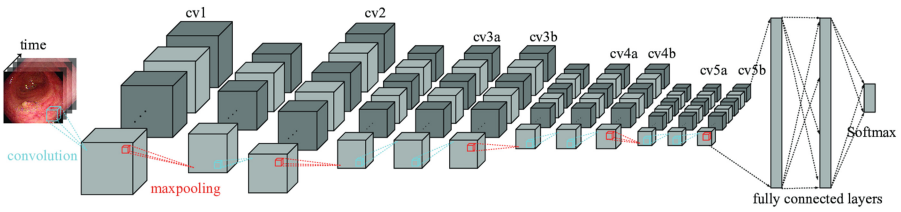


Fig. 2. Architecture of spatio-temporal classification for polyp detection. C3dNet extracts deep image spatial-temporal features via 3D convolutional and pooling procedures.

To extract and classify spatio-temporal image features for polyp detection, we use the 3D convolutional neural network (C3dNet) [8]. Figure 2 illustrates the C3dNet network architecture. The input for C3dNet is a set of successive 16

frames extracted from colonoscopic videos. We set all 3D convolutional filters as $3\times3\times3$ with $1\times1\times1$ stride. All 3D pooling layers are $2\times2\times2$ with $2\times2\times2$ stride, except for the first pooling layer which has kernel size of $1 \times 2 \times 2$. The output of C3dNet are the probability scores of two categories. If the output probability for positive category is larger than the criterion, polyp detection CAD system concludes that the input frames represent the scene where polyp exists. Note that before classification, we empirically search the best hyper-parameters of C3dNet to be optimized for polyp detection using the training dataset.

2.2 Two-Category Polyp Size Estimation

Our main purpose is to achieve the binary polyp size estimation (over 10 mm in diameter or not) from a 2D colonoscopic image $\mathcal{X} \in \mathbb{R}^{H\times W\times 3}$ of three channels with height H and width W. The straightforward estimation of polyp size s is defined as

$$\min \|s - s^*\|_2 \text{ w.r.t. } s = f(\mathcal{X}), \tag{1}$$

where $\|\cdot\|_2$ is the Euclidean norm, and s^* is the ground truth. This minimization problem is solved as regression that minimizes the square root error. However, this is an ill-posed problem since a 2D colonoscopic frame represents the appearance of a polyp on an image plane without available depth information. Therefore, we consider the function $f(\mathcal{X}, \boldsymbol{D^*})$ with the depth image $\boldsymbol{D}^* \in \mathbb{R}^{H\times W}$ that minimizes

$$\|s - s^*\|_2 \text{ w.r.t. } s = f(\mathcal{X}, \boldsymbol{D^*}). \tag{2}$$

We need annotated data with high precision to solve this minimization problem accurately. Note that polyp size annotation on images usually include small errors.

To make the polyp size estimation problem more practical and robust, we define the following relaxed minimization function with ground truth $s_B \in \{0, 1\}$ and $\mathcal{L}_0$-norm $\|\cdot\|_0$ as

$$\|f(\mathcal{X}, D^*) - s_B\|_0 \tag{3}$$

with respect to

$$f(\mathcal{X}, \boldsymbol{D^*}) = \begin{cases} 1, & \text{a polyp on an image is larger than 10 mm,} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Depth map information $\boldsymbol{D}^*$ is necessary in this definition although colonoscope device is not able to measure image depth values directly. In this relaxed form, we compute the depth image $D \in \mathbb{R}^{H\times W}$ that represents only relative depth information in an image, such like far and near. This type of depth cue $\boldsymbol{D}$ is not the physical distance from colonoscope to an object. Our depth images are obtained by adopting the unsupervised deep learning method of depth estimation from [2,3]. Using Depth or Disparity CNNs [2,3], we define a depth estimation function $g(\mathcal{X})$ that satisfies

$$\min \|g(\mathcal{X}) - \boldsymbol{D}^*/\|\boldsymbol{D}^*\|_{\mathrm{F}}\|_{\mathrm{F}}, \tag{5}$$

where $\|\cdot\|_{\mathrm{F}}$ is Frobenius norm, through unsupervised learning. This neural network need only colonoscopic videos for training, without ground truth of depth information (which is infeasible to acquire annotations for colonoscopic videos, if not entirely impossible).

Our proposed BsdNet shown in Fig. 3 intends to satisfy

$$\min \|f(\mathcal{X}, g(\mathcal{X})) - s_B\|_0 \text{ and } \min \|g(\mathcal{X}) - \boldsymbol{D}^*/\|\boldsymbol{D}^*\|_{\mathrm{F}}\|_{\mathrm{F}}. \tag{6}$$

The BseNet output the estimated size label $s \in \{0, 1\}$ for an input colonoscopic image. The right term of Eq. (6) is minimized by Depth CNN. The left term of Eq. (6) is minimized by RGB-D CNN shown in Fig. 4. The RGB-D CNN extracts RGB-D image features that minimizes the softmax loss function of two-category classification, that is, classification of polyps whether over or under 10 mm in diameter.

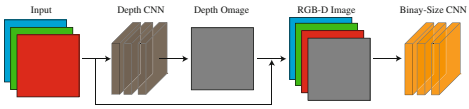


Fig. 3. Architecture of the binary polyp size estimation network (BseNet). BseNet first estimates the depth map from an RGB colonoscopic image by employing depth CNN. The estimated depth image is then combined with the input RGB channels to form an RGB-D image. BsdNet then classifies the newly composite RGB-D image into two categories: polyp over and under 10 mm in diameter.

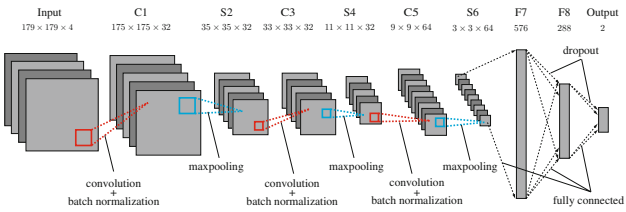


Fig. 4. Architecture of RGB-D CNN. Input is an RGB-D image of four channels.

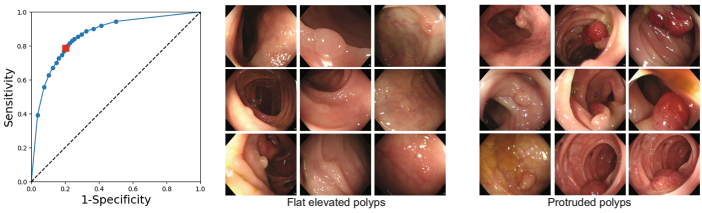
3 Experimental Results

Dataset: We construct a new dataset to validate our proposed polyp detection and binary size estimation method. We collect 73 colonoscopic videos, captured by CF-HQ290ZI (Olympus, Tokyo, Japan), with IRB approval. All frames of these videos are annotated by expert endoscopists. The total time of these videos

is about 16 h 37 min. The total video run time is 4 h 55 min (where 152 polyps are present or exist). These videos are captured under the different observation conditions of white light, narrow band imaging, and staining. Each frame is annotated and checked by two expert colonoscopists with experience over 5000 cases. Labels of pathological types, shape types, size (2, 3, . . . , 16 mm) and observation types are given.

3.1 Polyp Detection

We extract only the polyp frames that are captured under the white light condition. For non-polyp frames, we obtain images where polyps do not exist under several observation conditions. We divide these extracted frames into the training and testing datasets. The training dataset consist of polyp frames of 30 min 30 s and non-polyp frames of 24 min 12 s. The testing dataset represents of polyp frames of 18 min 1 s and non-polyp frames of 18 min 23 s. The training and testing datasets include different 102 and 50 polyps, respectively. Only training dataset is used for searching of optimal hyper-parameters of C3dNet with Adam optimizer. The testing dataset is used for validation of the classification accuracy of polyp and non-polyp frames. In both training and test datasets, colonoscopic images are rescaled into the resolution of 112×112 pixels. Therefore, the size of input data for c3dNet [8] is $112 \times 112 \times 16$. Figure 5 summarizes the validation results on the testing dataset.



(a) (b)

Fig. 5. Results for polyp detection. (a) Receiver Operating Characteristic (ROC) curve. (b) and (c) illustrate difficult and easy types, respectively, for detection.

Table 1. Results for each frame

	Accuracy for each frame
BseNet	79.2 %
CNN	77.5 %

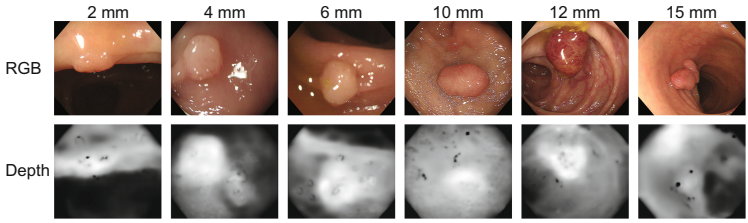


Fig. 6. The results of colonoscopic depth image estimation by unsupervised depth CNNs. White and black pixels represent near and far distances, respectively, from colonoscope.

Table 2. Results for each video clip

	RGB feature	RGB-D feature
Mean	55.3 %	73.7 %
LSTM	73.7 %	79.0

3.2 Polyp Size Estimation

Single Image Based Polyp Size Classification: We evaluated accuracy of the polyp size estimation as a frame classification problem for colonoscopic videos. We extracted 34,396 and 13,093 images of protrude polyps from 73 colonoscopic videos for training and test dataset for size estimation. The training and test datasets include different protrude polyps without duplication.

The training of BseNet is divided to two procedures. At the first training, we trained Depth CNN by using the polyp frames of 30 min 30 s in the previous subsection. For the second training, we estimated depth images of the training and test dataset and generated RGB-D images for training and test respectively. Figure 6 shows the depth images and original RGB images. At the second training, we trained RGB-D CNN with Adam for the generated RGB-D images. We evaluate the ratio of correct estimation as accuracy by using test dataset. For the comparison, we also trained RGB CNN and estimate polyp sizes by using the same training and test dataset of RGB-images. Table 1 summalizes the results of RGB-CNN and BseNet.

Video Clip Based Polyp Size Classification: We evaluated the polyp size estimation as a sequence classification problem with long-short term memory (LSTM) recurrent neural networks [9]. Given the per-frame predictions $P(\mathcal{X}_t) \in [0, 1]$ for over 10 mm size and per-frame penultimate feature response $\boldsymbol{F}(\mathcal{X}_t) \in \mathbb{R}^{288}$ of our size estimation of BseNet (F8 layer in Fig. 4) for a time sequence $t = 1, 2, \ldots,$, we build a sequence of feature vectors $\boldsymbol{f}_s = [P(\mathcal{X}_1), \boldsymbol{F}(\mathcal{X}_1)^\top, \ldots P(\mathcal{X}_n), \boldsymbol{F}(\mathcal{X}_n)^\top]^\top$ for LSTM classification. In our case, this results in a 289 length real valued vector for each frame of the sequence. We standardize all sequences to have zero-mean and std. dev. of one based on our training set. We furthermore limit the total length of a sequence to 1,000 by

either truncating the longer or padding the shorter polyp video clip feature vectors.

LSTM Model: We firstly use a stack of two LSTM layers consisting of 128 and 64 memory units each. The outputs from the second LSTM layer are then fed to two fully connected layers with 64 and 32 units, each employing batch normalization followed ReLU activations. A final fully connected layer predicts the polyp size from each sequence vector $\boldsymbol{f}_s$ with a *sigmoid* activation for binary classification.

Results are summarized in Table 2 and compared to using the average prediction value $|P(\mathcal{X}_t)|$ of all frames in the polyp sequence. As we can observe, both RGB and RGB-D cases experience an improved prediction accuracy using the LSTM model with the RGB-D model outperforming the model only based on color channels.

4 Discussion

When using the threshold criterion of 0.5 for polyp detection (see red square on Fig. 5), accuracy, sensitivity and specificity scores are 74.7%, 88.1% and 61.7%, respectively. The area under ROC curve (AUC) value is 0.83. In the current results, specificity is smaller than sensitivity, which implies the wider or broader varieties of patterns in the negative class of non-polyp frames for polyp detections. In these experiments, the detection rate of flat elevated polyp as shown in Fig. 5(b) is smaller than the detection rate of protruded polyps, demonstrated in Fig. 5(c).

The experimental results for size estimations show that our proposed BseNet (using RGB+D) achieves 79.2% accuracy for binary polyp-size classification that is about 2% larger than the accuracy of CNN (only using RGB). This results imply the validity of relaxed form of size estimation. We also combine the image feature extraction by BseNet and classification of short video clips using a long short-term memory (LSTM) network. The results of LSTM classifications also show that RGB+D features that extracted by BseNet achieves 5.3% higher accuracy than RGB features alone. These results show the validity of RGB-D features extracted by BseNet.

5 Conclusions

We formulated the relaxed form of polyp size estimation from colonoscopic video as the binary classification problem and solve it by proposing the new deep learning-based architecture: BseNet towards automated colonoscopy diagnosis. BseNet estimates the depth map image from an input colonoscopic RGB image using upsupervised deep learning, and integrates RGB with the computed depth information to produce four-channel RGB-D imagery data. This RGB-D data is subsequently encoded by BseNet to extract deep RGB-D image features and facilitate the size classification into two categories: under and over 10 mm polyps.

Our experimental results show the validity of the relaxed form of the size estimation and the promising performance of the proposed BseNet.

This research was partially supported by AMED Research Grant (18hs0110006 h0002, 18hk0102034h0103), and JSPS KAKENHI (26108006, 17H00867, 17K20 099).

References

1. Winawer, S.: Guidelines for colonoscopy surveillance after polypectomy: a consensus update by the us multi-society task force on colorectal cancer and the american cancer society. Gastroenterology **130**, 1872–1885 (2006)
2. Mayer, N., et al.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the CVPR (2016)
3. Zhou, T., Brown, M., Snavely, N., Lowe, D.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the CVPR (2017)
4. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2011)
5. Visentini-Scarzanella, M., Stoyanov, D., Yang, G.Z.: Metric depth recovery from monocular images using shape-from-shading and specularities. In: Proceedings of the IEEE ICIP (2012)
6. Bernal, J., Sánchez, J., Vilariño, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognit. **45**(9), 3166–3182 (2012)
7. Bernal, J.: Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. IEEE Trans. Med. Imaging **36**(6), 1231–1249 (2017)
8. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the ICCV (2015)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)