2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)
Venice, Italy, April 8-11, 2019

POLYP TRACKING IN VIDEO COLONOSCOPY USING OPTICAL FLOW WITH AN ON-THE-FLY TRAINED CNN

He Zheng[1,2,3] *Hanbo Chen*[1] *Junzhou Huang*[1] *Xuzhi Li*[2,3] *Xiao Han*[1] *Jianhua Yao*[1]

[1]Tencent AI Lab; [2]University of Chinese Academy of Sciences, Beijing, China;
[3]Key Laboratory of Space Utilization,
Technology and Engineering Center for space Utilization, Chinese Academy of Sciences, Beijing, China

ABSTRACT

Colonoscopy has been widely applied as a common practice to inspect the inside of large bowel for colon cancer screening. However, missing polyps in such procedure could happen and thus preventing early disease detection and treatment. In this paper, we propose an algorithm for automatic polyp detection and localization in colonoscopy video. The method initially detects and localizes polyps based on single frame object detection or segmentation network such as U-Net. Then it utilizes optical flow to track polyps and fuse temporal information. To overcome tracking failure caused by motion effects, a motion regression model and an efficient on-the-fly trained CNN have been deployed. The proposed algorithm achieves the highest scores in both polyp detection task and polyp localization task in the MICCAI 2018 Endoscopic Vision Challenge on ”Gastrointestinal Image Analysis”.

***Index Terms*—** polyp detection, polyp localization, video colonoscopy, tracking, on-the-fly trained cnn, optical flow, motion regression

1. INTRODUCTION

Colorectal cancer (CRC) is one of the most prevalent causes of cancer death worldwide. Currently, the standard approach to reduce CRC-related mortality is through efficient and regular colon screening to search for polyps. As today’s common practice, colonoscopy has been widely applied for colon cancer screening. During a colonoscopy exploration, clinicians inspect the intestinal wall to detect polyps [1, 2]. However, missing polyps in such procedure could happen and thus miss the chance for early disease detection and treatment.

To reduce the risk of miss-diagnosis and ease doctors’ burden, Computer-Aided Detection (CAD) methods for polyp detection during colonoscopy have been investigated [3]. Compared to the CAD tasks in the non-medical domain, automated polyp detection is technically challenging in practice since 1) the same type of polyps can vary significantly in size, color and texture, and 2) many polyps do not stand out clearly from the surrounding mucosa. Previous CAD systems usually extracted handcrafted features from polyp images, such as their color, texture, shape or appearance, or the combination of these features, and trained a classifier to distinguish the polyp from the surroundings. More recent studies adopted the convolutional neural network (CNN) architecture for polyp detection [4, 5, 6] given CNN’s great success in automatically learning representations of target objects in computer vision tasks.

Notably, most of the existing CNN approaches detect polyps on each frame independently. This may result in jittering effect due to minor intensity fluctuation between frames. In [7], the authors showed that by utilizing a tracker to refine the detection results on each frame generated by CNN can significantly reduce jittering. However, the method proposed in [7] required accurate detection of the polyp in the initial frame which is difficult to achieve.

The major challenge in training a CNN model to accurately detect polyp is the heterogeneous appearance of polyps between video frames. However, in our observation, the polyp characteristics are relatively consistent for each patient, and the motion change between adjacent frames is relatively small shown in Fig. 3. Inspired by such observation, in this work, we propose an optical flow model combined with an on-the-fly trained CNN (OptCNN) model and a spatial voting algorithm to refine the detection results given by a single frame polyp detector.

2. METHOD

Overview of our proposed method is shown in Fig. 1. A CNN model is first trained to detect and segment polyp in each video frame (2.1). Once a polyp is detected, the center of the polyp is computed and traced through following frames until stop criteria is met (2.2). During tracing, optical flow is utilized to trace easier cases (2.2.1) and OptCNN is used to process harder ones (2.2.3). If a frame does not contain any polyp center seed, the frame will be regarded as a negative frame (no polyp appears). If there are multiple polyp seeds in a frame (some are traced from previous frames), a spatial voting algorithm is run and the most confident center is kept as the detection while others are eliminated (2.3).

978-1-5386-3640-4/19/$31.00 ©2019 IEEE

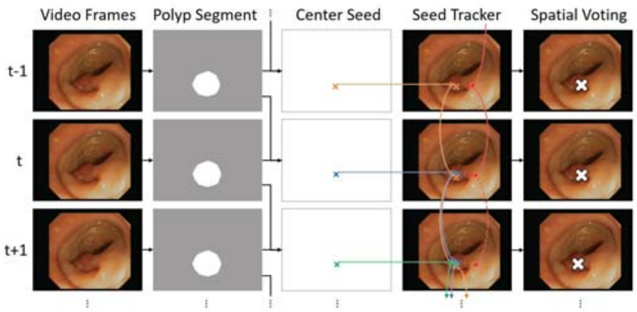

Fig. 1. The overview of the proposed algorithm.

2.1. Single Frame Polyp Detector

We adopt U-net to identify polyp areas on each single frame. U-net is an end-to-end CNN that can take an image as input and output segmentations of object of interest in the image [8]. The most distinct feature of this network is the skip connection between encoder and decoder. This network architecture has been widely applied in different CAD systems and achieved great success.

We train an U-net to predict a probability map with 1 indicates polyp and 0 indicates none-polyp area. To reduce jittering effect, we average the map generated from current frame with the one generated from the previous frame. The pixels with probability larger than 0.5 is then taken as foreground. Then connected components are extracted and filtered to only keep the one with round shape and large size. In brief, we first conduct erosion to smooth the spur in the segmentation boundary. Then we compute connected components and rank them by size and roundness separately. Only when the largest area and the most ellipse-like shape area are the same component, the center of this component will be computed and taken as a new polyp center seed. Otherwise, we will rely on the tracking algorithm introduced later to localize the polyp center in this frame.

2.2. Polyp Tracker

2.2.1. Optical Flow

Optical flow is the pattern of apparent motion of image objects between two consecutive video frames. It is a 2D vector field where each vector is a displacement vector showing the flow of points from the first frame to the second [9].

Optical flow is based on two assumptions:

1. The pixel intensities of an object do not change between consecutive frames.

2. Neighboring pixels have similar motion.

Giving the polyp center (x, y) in frame t, we use the method proposed in [10] to trace its location in the next frame. However,a blurred image or image artifact may cause the above two assumptions not satisfied, and thus results in an early stop for the optical flow method. To decide whether to continue tracking, we use a more robust deep learning based regression and classification method to evaluate and conduct further tracking when the optical flow tracking stops.

2.2.2. Motion Regression Model

A regression model is designed to estimate the motion of a polyp center based on the sequence of its movements in the previous frames. Given $\Delta P_t = P_t - P_{t-1}$ represents the motion of polyp center in frame t. Using a training sample $\{\mathbf{x}_i, y_i\}$, the goal is to obtain an estimator approximation $\hat{F}(\mathbf{x})$, of the function $F(\mathbf{x})$ mapping $\mathbf{x}$ to y, that minimizes the expected value of some specified loss function $L(y, F(\mathbf{x}))$ over the joint distribution of all $(\mathbf{x}, y)$-values [11].

$$F^* = \arg\min_F E_{\mathbf{x}}[E_y(L(y, F(\mathbf{x})))|\mathbf{x}] \qquad (1)$$

Where $\mathbf{x} = [\Delta P_{t-3}, \Delta P_{t-2}, \Delta P_{t-1}]$ is a three dimension vector. $y = \Delta P_t$. We then added P_{t-1} with ΔP_t to estimate its new location. If the new location is inside the video frame, a CNN model will then be used to decide whether it is a polyp center (continue tracking) or not (stop tracking).

2.2.3. On-the-fly Trained CNN

Based on the observation that the appearance of a polyp stays consistent between frames, we propose an on-the-fly trained CNN-based framework to classify polyp from background and decide whether the tracking estimated by regression model should stop or not. This method is inspired by [12] and outlined in Fig. 2.

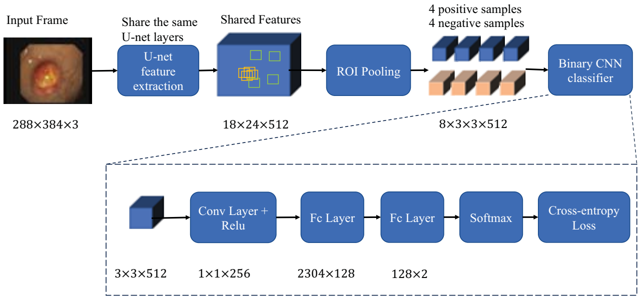

Fig. 2. On-the-fly trained CNN.

Given a frame t, four positive samples (jaccard overlap to polyp centered box larger than 0.7) and four negative samples (overlap smaller than 0.1) were generated from its previous frame $t-1$. A CNN classifier is fine-tuned (trained on-the-fly) based on these 8 samples to classify polyp from background. Then for the estimated center location on the current frame, the classifier will predict whether it is polyp or not. Tracking will stop if the classifier classify it as background.

For speed reason, the high-level feature map previously generated by U-net's encoder when conducting segmentation task is directly used as the input of a small network which is trained to predict polyps. In our experiment, the input image size is $288 \times 384 \times 3$. The dimension of feature map generated by U-Net is $18 \times 24 \times 512$ which is 16 times smaller in size than raw image. Thus when extracting the feature map of region of interests (ROIs), we reduced the size of ROI by 16 times and directly crop the feature map to the corresponding size. For simplicity, we fix the width and the height of the ROI to 48×48. 3×3 area is then cropped on feature map and fed into a $1 \times 1 \times 256$ convolution layers followed by two fully connected layers and a softmax layer. We use the cross-entropy loss when fine tuning this network.

2.2.4. Tracking Stop Criteria

The tracking of a polyp center stops when the estimated location of seed is out of frame or the on-the-fly trained CNN classifier classifies the seed as background. Since the same polyp could be traced by multiple seeds generated from different frames, to save computation time and prevent unnecessary tracking, we stop tracking when the seed has been tracked in more than 10 frames. To reduce the error caused by on-the-fly trained CNN, we stop tracking when the optical flow method failed to track in 3 successive frames.

2.3. Spatial Voting

After tracking, one frame may contain multiple seeds. One way is to average all seeds as the polyp center. However, some seeds could be outliers and thus affect the results. Based on the observation that correct polyp center will concentrate in a small area, we proposed a spatial voting algorithm to eliminate outliers. First, the adjacent seeds with Euclidean distance smaller than threshold were connected. Then the connected component was computed and the center of the largest component was taken as the final polyp center.

3. EXPERIMENTS

3.1. Datasets

In this paper, we use the training datasets of the sub-challenging "Gastrointestinal Image Analysis" in MICCAI 2018 Endoscopic Vision Challenge, which is part of CVC-VideoClinicDB database [1, 2]. This database is composed of video sequences extracted from routine clinical exams at Hospital Clinic of Barcelona, Spain and aims to cover all different scenarios that a given CAD system should face.

The data we applied in this paper is composes of 18 different video sequences. In each video, at least one polyp has been observed in part of the sequence. Clinical experts have reviewed the data frame by frame and annotated the area of polyps with eclipse when they appear. Notably, only 1 polyp will be annotated in each frame. To train and evaluate our proposed method, we split the dataset into training (14 videos) and testing set (4 videos).

3.2. Performance Metrics

Two different tasks have been evaluated in our experiments. 1) **Polyp detection**: we predict whether there is a polyp in each frame. 2) **Polyp localization**: when there is a polyp in the frame, predict the center of the polyp. For each task, we evaluate 4 critical metrics: precision, recall, F1-score $(2 \times precision \times recall/(precision + recall)$, F2-score $(5 \times precision \times recall/(4 \times precision + recall))$. Specifically, for polyp center prediction task, only when the predicted center is inside the annotation mask which is provided by the dataset, it will be taken as true positive. Otherwise, it will be counted as a false positive. And when an annotated mask is not touched by predicted centers, a false negative instance occurs.

3.3. Results

In this section, we present the performance achieved by our method on the testing data. For comparison, we also compute results by using U-Net only, U-Net with temporal fusion (U-net+fuse), and by using optical flow only when tracking center seeds. Fig.3 shows an example of tracking from the two methods, optical flow only and the proposed method. At Frame 4 Optical flow stops tracking but the CNN-based method keeps tracking. At Frame $f+2$, the proposed method stops tracking based on the termination strategy. As we can see, the proposed method performs reasonably well in detecting and tracking polyps even when there is considerable motion artifact in the image.

Table 1 shows the quantification results of the polyp detection task. We can see that post-processing of utilizing temporal information could slightly improve the performance by reducing jittering effect. By using optical flow to track detected polyp seeds through time can significantly reduce the number of undetected polyps and increase recall. At meanwhile, this may also increase the chance of false positive and thus reduce precision. However, the overall performance has been improved and significant improvement on F1 and F2 has been achieved. By adding OptCNN as we proposed can further improve the performance.

Table 1. Results of the videos for testing of polyp detection.

Methods	Precision	Recall	F1	F2
U-net	89.69	75.86	82.20	78.28
U-net+fuse	89.66	76.01	82.27	78.40
Optical flow	86.15	94.98	90.35	93.07
Proposed	84.58	97.29	90.49	94.45

Table 2 shows the performance of the polyp localization task. Similar to the polyp detection task. Proposed method performed the best in catching most of polyps and yield the highest recall and F1-score. However, due to the increased chance of detecting False Negatives, there is decreasing in precision and F1-score.

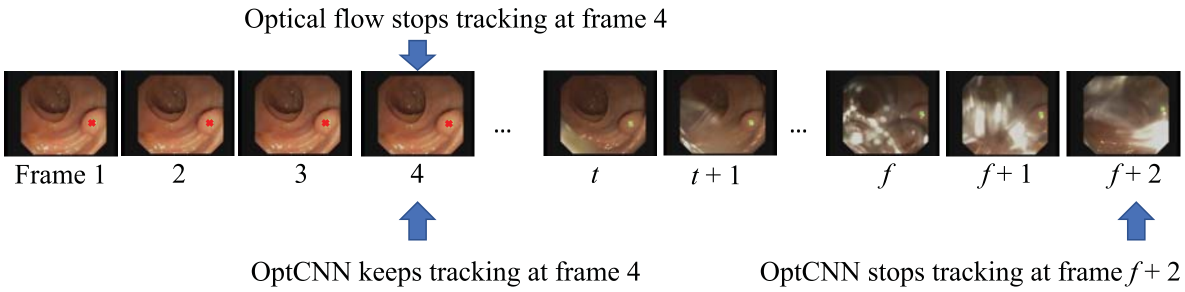
Optical flow stops tracking at frame 4
Frame 1
2
3
4
...
t
t + 1
...
f
f + 1
f + 2
OptCNN keeps tracking at frame 4
OptCNN stops tracking at frame f + 2

Fig. 3. An example of tracking 1 center seed. Red dots are the tracking generated by optical flow. Green dots are the tracking failed by optical flow and generated by OptCNN instead.

The Polyp Tracker contributes the most in the pipeline because we utlize the information between the frames. Both optical flow and OptCNN can achieve significant improvement on F1 and F2. The OptCNN as we proposed can further improve the performance on Recall.

Table 2. Results of the videos for testing of polyp localization.

Methods	Precision	Recall	F1	F2
U-net	81.69	71.97	76.52	73.72
U-net+fuse	81.59	72.11	76.56	73.82
Optical flow	78.11	93.66	85.18	90.07
proposed	74.28	96.39	83.90	90.97

4. CONCLUSION

In conclusion, the proposed pipeline performs reasonably well in detecting and localizing polyps in colonoscopy videos. It achieved the highest scores in both polyp detection task and polyp localization task in the MICCAI 2018 Endoscopic Vision Challenge on ”Gastrointestinal Image Analysis”. However, how to balance between false positive and false negative discoveries with the proposed method is still an open question. With high recall and reasonably precision performance, we believe the proposed method can be clinically used to alert colonoscopy operator when polyp is detected and thus help to reduce the rate of under-diagnosis.

5. REFERENCES

[1] Quentin Angermann, et al., “Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis,” in *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*, pp. 29–41. Springer, 2017.

[2] Jorge Bernal, et al., “Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases,” in *CARS*, 2018.

[3] Yuichi Mori, et al., “Computer-aided diagnosis for colonoscopy,” *Endoscopy*, vol. 49, no. 08, pp. 813–819, 2017.

[4] Zijie Yuan, et al., “Automatic polyp detection in colonoscopy videos,” in *Medical Imaging 2017: Image Processing*. International Society for Optics and Photonics, 2017, vol. 10133, p. 101332K.

[5] Nima Tajbakhsh, et al., “Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks,” in *ISBI*. IEEE, 2015, pp. 79–83.

[6] Jorge Bernal, et al., “Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge,” *IEEE transactions on medical imaging*, vol. 36, no. 6, pp. 1231–1249, 2017.

[7] Ruikai Zhang, et al., “Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker,” *Pattern Recognition*, 2018.

[8] Olaf Ronneberger, et al., “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[9] Andrés Bruhn, et al., “Lucas/kanade meets horn/schunck: Combining local and global optic flow methods,” *International journal of computer vision*, vol. 61, no. 3, pp. 211–231, 2005.

[10] Bruce D Lucas, et al., “An iterative image registration technique with an application to stereo vision,” 1981.

[11] Jerome H Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.

[12] Qi Chu, et al., “Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism,” in *ICCV*, 2017, pp. 4846–4855.