Deep Learning using temporal information for automatic polyp detection in videos

Adrian Krenzer[1], Philipp Sodmann[2], Nico Hasler[1] and Frank Puppe[1]

[1]*Department of Artificial Intelligence and Knowledge Systems, University of Würzburg, Germany*
[2]*Gastroenterology department of the University Hospital of Würzburg, University of Würzburg, Germany*

Abstract

Previous research in the field of endoscopic computer vision has mainly focused on the detection of polyps using single images, but not videos or streams of images. The Endoscopic computer vision challenges 2.0 (EndoCV 2.0) is designed specifically to use streams of image sequences for the detection of polyps. In this paper, we describe our approach based on Gong et al. [1] by leveraging deep convolutional neural networks (CNNs) combined with temporal information to improve upon existing solutions for polyp detection. We demonstrate a detection system that combines similar ROI features across multiple frames with temporal attention to predict the final polyp detections for an emerging frame. For evaluation, we compare our approach to two classical image detection algorithms on a validation set based on training data provided by the challenge. The first one is a Single Shot Detector (SSD) called "YOLOv3", and the second one is a two-step region proposal-based CNN called "Faster R-CNN". To minimize the generalization error, we apply data augmentation and add additional open-source data for our training.

Keywords

Machine learning, Deep learning, Endoscopy, Automation, Video object detection, Attention

1. Introduction

The second leading cause of cancer-related deaths worldwide is Colorectal cancer (CRC) [2]. An excellent method to prevent CRC is to detect pre-cancerous lesions (colorectal polyps) of the disease as early as possible, using a colonoscopy. During a colonoscopy, a long flexible tube that is inserted through the rectum into the colon. The end of the tube has a small camera, allowing the physician to examine the colon thoroughly [1]. Computer science researchers are developing new methods to support physicians with this procedure. Polyp detection using computers is called *computer-aided detection (CAD)*. This process of polyp detection has already been subject to numerous publications.

However, these published solutions mostly focus on detection on still images [3]. Therefore, most of the published algorithms do not consider temporal dependencies and do compare themselves on benchmarks which do not consider temporal connections. To predict the final polyp detections for an emerging frame, our approach based on Gong et al. [1] utilizes temporal dependencies by combining similar ROI features across successive frames with temporal attention. Nevertheless, there are already some approaches in the literature addressing temporal dependency in polyp detection: In Itoh et al. [4], temporal information is included through a *3D-ResNet*. The 3D ResNet is thereby combining present and future frames for the detection of a new frame.

Furthermore, Qadir et al. [5] work with a traditional localization model, such as SSD [6] or Faster R-CNN [7], and post-process the output with an *FP Reduction Unit*. This approach considers the area of the generated bounding boxes over the 7 preceding and following frames and identifies and adjusts the outliers. The use of future frames causes a small delay, however, the actual calculation of the *FP Reduction Unit* is fast. A second promising method by Qadir et al. uses a two-step process which aims to decrease the proportion of false predictions. Furthermore, the CNN that flags several regions of interest (ROIs) for classification. The marked ROIs are then compared with subsequent frames and their corresponding ROIs and classified into true positives and false positives. The underlying assumption here is that each frame in a video is similar to its adjacent frames [5].

Xu et al. [8] designed a 2D CNN detector, which takes the spatiotemporal information into account and uses an ISTM network to improve its polyp detection efficiency while maintaining real-time speed. The model was trained on custom data. In addition, there is another approach which includes the temporal dependencies via post-processing. This approach uses fast image detection algorithms like YOLO and, afterwards, combines these predictions with an efficient real-time post-processing technic. This post-processing technique includes the predictions of polyps detected in past frames for future

4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) in conjunction with the 19th IEEE International Symposium on Biomedical Imaging ISBI2022, March 28th, 2022, IC Royal Bengal, Kolkata, India

adrian.krenzer@uni-wuerzburg.de (A. Krenzer)

[1] https://www.mayoclinic.org/tests-procedures/colonoscopy/about/pac-20393569

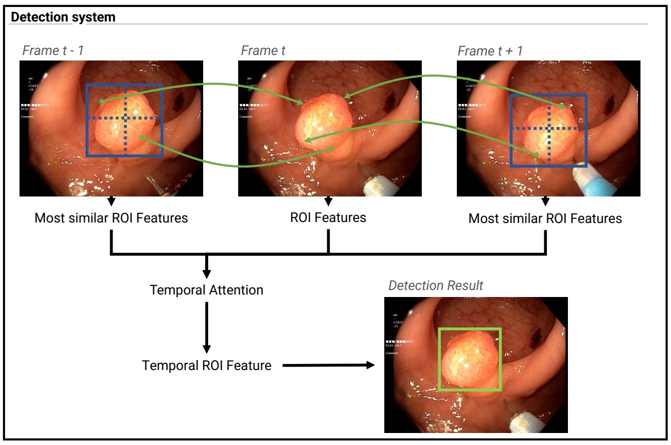

Figure 1: Overview of the polyp detection approach. t denotes the current frame for the detection. t - 1 denotes the frame before frame t and t+1 the frame after frame t. The ROIs are aligned through temporal attention for different frames. This figure is adopted from Gong et al. [1]

.

detections [9]. Taking these ideas forward, we implemented a polyp-detection model using the "ROI-Align Module" of Gong et al. [1] This allows the neural network to attend to information in previous frames and to combine ROI features from different frames for new predictions.

2. Data

To train the model, we used two public available datasets in addition to the challenge dataset:

- Kvasir-SEG [10]: 1000 polyp frames are included in the data collection, along with 1071 masks and bounding boxes. The sizes range from 332×487 pixels to 1920×1072 pixels. Gastroenterologists at Norway's *Vestre Viken Health Trust* confirmed the annotations. The majority of the frames show basic information on the left side, while others have a black box in the lower-left corner that contains data from ScopeGuide's endoscope position marking probe (Olympus). The data is available in the Kvasir-SEG repository[2].

- SUN Colonoscopy Video Database [11]: This dataset comprises 49,136 polyp frames from 100 distinct polyps, all of which are thoroughly documented. These frames were taken at Showa University Northern Yokohama and annotated by Showa University's specialist endoscopists. There are also 109,554 non-polyp frames present. The frames have a resolution of 1240×1080 pixels. The data is available in the SUN Colonoscopy Video repository[3].

- PolypGen2.0 (Polyp Generalization) [12, 13, 14]: This dataset is one of the two sets from the challenge and an extended version of the datasets from the 2020 and 2021 challenges. Both sub-challenges provide multi-center and diverse population datasets with tasks for both detection and segmentation, but the emphasis is on evaluating algorithm generalizability. The goal was to incorporate additional sequence/video data as well as multimodal data from various sites. PolyGen2.0 consists of 46 sequences with a total of 3290 images. All frames have a resolution of 1920×1080 pixels.

We split the PolyGen2.0 dataset into training and validation. For this purpose, 20 random sequences were assigned to validation (1366 images) and the rest to training (1924 images). The resulting validation set was used for all training steps.

3. Methods

In this section, we illustrate our approaches for the EndoCV2022 challenge, depicted in figure 1. All our models are trained on a NVIDIA QUADRO RTX 8000. After exploring the data, we decided to choose an algorithm which includes temporal information for the challenge, since the test data provided includes entire videos rather

[2] https://datasets.simula.no/kvasir-seg/

[3] http://sundatabase.org/

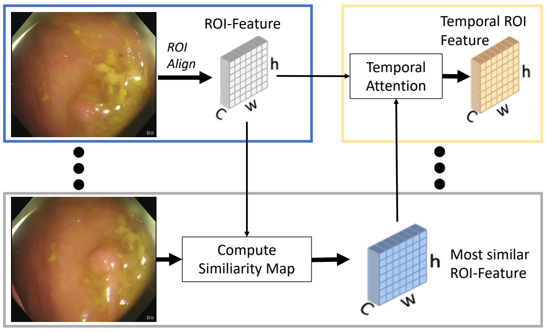

Figure 2: This figure illustrates temporal ROI align design and how its similarity map aggregation and temporal attention are used to compute the temporal ROI feature. This figure is adopted from Gong et al. [1]

than just images. The model is based on Gong et al. [1] and will be explained in the following.

Most state-of-the-art single-frame object detectors use the paradigm of region-based detection. When these detectors are used directly for video object detection (VID), object appearances in videos such as motion blur, video defocus, and object occlusions can degrade detection accuracy. These are frequent problems in endoscopy videos, which make the detection of polyps more difficult. Therefore, the main challenge is to design a method that can utilize the temporal redundancy of the information efficiently for the same object instance in a sequence of images or videos. To extract ROI features, most region-based detectors use ROI Align. However, ROI Align only uses the current frame feature map to extract features for current frame proposals, resulting in ROI features that lack the temporal information of the same object instance in the video. Using feature maps of other frames to perform ROI Align for the current frame proposals is a straightforward and clear technique for using temporal information. However, since the exact placement of the current frame proposals in other frame feature maps is unknown, the basic solution is ineffective.

Temporal ROI Align, on the other hand, defines a target frame as a frame in which the final prediction is made in real-time. In figure 2 the temporal ROI algin process is illustrated. Temporal ROI algin also allows the target frame to have multiple support frames, which are used to refine the features of the target frame. To achieve this refinement, the proposed operator selects the most comparable ROI features from the feature maps of the available support frames. The temporally redundant information of the same object instance in a video is contained in the extracted most comparable ROI characteristics. The main target now is to effectively capture diverse ROI features. Average is inefficient, because a polyp may seem blurry in some frames and clear in others. It is self-evident that the ROI characteristics of clear object instances should take precedence over the features of blurry instances in aggregate. To aggregate the ROI characteristics and the most comparable ROI features, multi-temporal attention blocks are used to perform the temporal feature aggregation. A major advantage of Temporal ROI Align is that it can extract the object features from support frames even when a polyp is partially occluded in the target frame. Therefore, the visible parts are dominant and features at these locations can still get enhanced.

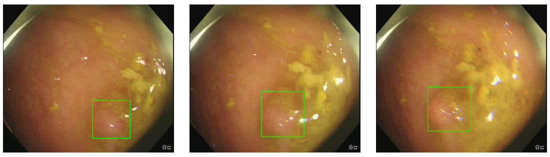

Figure 3: This figure shows a sequence of detections results with our algorithm on the test dataset provided by the challenge. Time is in this sequence running from the left side image to the right side while the polyp is moving to the left.

For our approach, the nerual network is trained for 10 epochs on our full dataset and then finetuned for 3 epochs on the challenge dataset. We choose the stochastic gradient descent (SGD) optimizer with a learning rate of 0.01, momentum of 0.9, and a weight decay of 0.0001. Additionally, we use a linear training warm-up schedule for 1 epoch. To enhance the generalization capabilities of our model, we use the following augmentation-schema: We applied a probability of 0.3 for upward and downward flips and a vertical flipping probability of 0.5. In addition, we rescaled the image with a probability of 0.64. We also use a translation along the horizontal axis with a probability of 0.5.

4. Results

In this section, we describe our results of the EndoCV2022 challenge. We highlight the performance of our approach and compare it to two classic benchmarking algorithms. One is an SSD algorithm called YOLOv3 [15] and the other is the ROI Proposal algorithm called Faster RCNN [16]. We trained both algorithms on the same data. For the validation, we create a validation set. The validation set consists of 20 sequences randomly chosen from the provided data (no additional data is included). We test the detection-created validation set. To enable the comparison of our results with the other participants of the challenge we do also declare our final scores: Score(mAP) 13.12 % and score(mAP50) 27.05 % are our final detection scores on the second round of the challenge evaluation.

Table 1 shows our results on our created validation set for the detection task where YOLOv3 is a benchmark

SSD algorithm, Faster R-CNN is the FASTER R-CNN algorithm with ResNet-101 backbone. For the evaluation, we report the F1-score. The F1-score describes the harmonic mean of precision and recall as shown in the following equations:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * TP}{2 * TP + FP + FN}$$

We count an annotation as true positive (TP) if the boxes of our prediction and the boxes from the ground truth overlap at least 50%. Additionally, we display the mean average precision (mAP) and the mAP50 with a minimum IoU of 0.5 [17]. The mAP is calculated by the integral of the area under the precision-recall curve. Thereby, all predicted boxen are first ranked by their confidence value given by the polyp detection system. Then we computed precision and recall for different thresholds of these confidence values. When reducing the confidence threshold recall increases and precision decreases. This results in a precision-recall curve. Finally, for this precision-recall curve, the area under the curve is measured. This results in the mAP.

Table 1 shows that our approach is outperforming classical benchmarks on our validation data; this is mostly due to our temporal dependencies included in the algorithm which are not included in the Faster-RCNN approach. Notably, SSD algorithms like YOLOv3 are still 20 FPS faster than our approach in detecting single images. Nevertheless, our approach yield a huge recall increase of 9.5 % compared to the fast YOLOv3. We do especially emphasize this as recall is one of the most important metrics in real clinical use. As it is more important to find a missing polyp than to have additional false positiv detections. Figure 3 shows a sequence of detections results with our algorithm on the test dataset provided by the challenge. Furthermore, figure 4 shows a qualitative comparison of the three detection algorithms. We can see that all algorithms are detecting the polyp. Nevertheless, Yolov3 and Faster-RCNN are distracted by light reflections and therefore also draw wrong detections. Through temporal ROI align, our approach can incorporate the detections from previous frames and therefore does not get distracted by the light reflections.

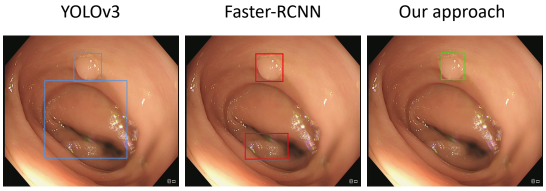

Figure 4: This figure shows a qualitative comparison of the three detection algorithms.

Table 1
Evaluation results of our validation split. We compare our approach based on Gong et al. [1] to two different polyp detection baselines on the same validation split from the challenge. Precision, Recall, F1, and mAP are given in %, and the speed is given in FPS.

	YOLOv3	Faster-RCNN	Our approach
mAP	13.8	14.2	**18.8**
mAP50	27.5	28.9	**32.8**
Precision	32.2	**34.5**	32.4
Recall	30.1	32.4	**39.6**
F1	31.1	33.4	**35.6**
Speed	**44**	15	24

5. Discussion

In this section, we like to discuss two main points: First, the limitations of our approach, and second how to use our approach in clinical useful settings. The first limitation is the current speed of our system. With an inference performance of 24 FPS, the algorithm is not capable of detecting every image with an endoscopy processor processing at 30 FPS. This can be mitigated by pruning and quantization-aware retraining. This on the other hand reduces the accuracy of the algorithm. Additionally, in the literature, a lot of benchmarking scores on still polyp images are already exceeding 80 % F1 score [18, 19]. Nevertheless, those are not directly comparable with our evaluation as they are using different data sets and do not include sequences of images.

The second and most drastic issue is that the system in its current form only works with video data and not a real-time stream of videos due to the dependencies in the algorithm, including preceding and future frames in the prediction. This issue may be solved by changing the algorithm to only use the preceding frames. In its current form, the algorithm can be used to evaluate endoscopies after they are completed or to detect polyps with wireless capsule endoscopy (WCE).

6. Conclusion

Overall, we demonstrate our approach to the Endoscopic computer vision challenges 2.0. We show a detection system that combines similar ROI Features across frames with temporal attention to create the final for polyp detections for a new emerging frame. The system thereby uses present, past, and future features on the temporal axis to create new polyp localizations. We show that the system exceeds classical benchmarks algorithms based on individual frames on our validation data from the challenge.

7. Compliance with ethical standards

This research study was conducted retrospectively using human subject data made available in open access [10, 11, 12, 13, 14]. Ethical approval was not required as confirmed by the license attached with the open access data.

8. Acknowledgments

This research is supported using public funding from Interdisziplinäres Zentrum für Klinische Forschung (IZKF) from the University of Würzburg.

References

[1] T. Gong, K. Chen, X. Wang, Q. Chu, F. Zhu, D. Lin, N. Yu, H. Feng, Temporal roi align for video object recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 1442–1450.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians 68 (2018) 394–424. URL: https://doi.org/10.3322/caac.21492. doi:10.3322/caac.21492.

[3] A. Krenzer, A. Hekalo, F. Puppe, Endoscopic detection and segmentation of gastroenterological diseases with deep convolutional neural networks., in: EndoCV@ ISBI, 2020, pp. 58–63.

[4] H. Itoh, H. Roth, M. Oda, M. Misawa, Y. Mori, S.-E. Kudo, K. Mori, Stable polyp-scene classification via subsampling and residual learning from an imbalanced large dataset, Healthcare Technology Letters 6 (2019) 237–242. URL: https://doi.org/10.1049/htl.2019.0079. doi:10.1049/htl.2019.0079.

[5] H. A. Qadir, I. Balasingham, J. Solhusvik, J. Bergsland, L. Aabakken, Y. Shin, Improving automatic polyp detection using CNN by exploiting temporal dependency in colonoscopy video, IEEE Journal of Biomedical and Health Informatics 24 (2020) 180–193. URL: https://doi.org/10.1109/jbhi.2019.2907434. doi:10.1109/jbhi.2019.2907434.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, ArXiv abs/1512.02325 (2016).

[7] S. Ren, K. He, R. Girshick, J. Sun, Faster r-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 1137–1149. URL: https://doi.org/10.1109/tpami.2016.2577031. doi:10.1109/tpami.2016.2577031.

[8] X. Liu, X. Guo, Y. Liu, Y. Yuan, Consolidated domain adaptive detection and localization framework for cross-device colonoscopic images, Medical image analysis 71 (2021) 102052.

[9] A. Krenzer, M. Banck, K. Makowski, A. Hekalo, D. Fitting, J. Troya, B. Sudarevic, W. G. Zoller, A. Hann, F. Puppe, A real-time polyp detection system with clinical application in colonoscopy using deep convolutional neural networks (2022).

[10] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: International Conference on Multimedia Modeling, Springer, 2020, pp. 451–462.

[11] M. Misawa, S.-e. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida, et al., Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video), Gastrointestinal Endoscopy 93 (2021) 960–967.

[12] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler, et al., Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv preprint arXiv:2106.04463 (2021). doi:10.48550/arXiv.2106.04463.

[13] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. M. et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical Image Analysis 70 (2021) 102002. URL: https://www.sciencedirect.com/science/article/pii/S1361841521000487. doi:https://doi.org/10.1016/j.media.2021.102002.

[14] S. Ali, N. Ghatwary, D. Jha, E. Isik-Polat, G. Polat, C. Yang, W. Li, A. Galdran, M.-Á. G. Ballester, V. Thambawita, et al., Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge, arXiv preprint arXiv:2202.12031 (2022). doi:10.48550/arXiv.2202.12031.

[15] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018).

[16] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona,

D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[18] D. Wang, N. Zhang, X. Sun, P. Zhang, C. Zhang, Y. Cao, B. Liu, Afp-net: Realtime anchor-free polyp detection in colonoscopy, in: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2019, pp. 636–643.

[19] X. Mo, K. Tao, Q. Wang, G. Wang, An efficient approach for polyps detection in endoscopic videos based on faster r-cnn, in: 2018 24th international conference on pattern recognition (ICPR), IEEE, 2018, pp. 3929–3934.