Semi-supervised video object segmentation as annotation tool for endoscopic video

Tom Eelbode[1], Paul Suetens[1], Raf Bisschops[2], and Frederik Maes[1]

[1] ESAT, Center for Processing Speech and Images, KU Leuven, Belgium
[2] Gastroenterology and hepatology, KU Leuven, Belgium

1 Introduction

Deep learning techniques have recently been reported to be highly performant for the detection and segmentation of endoscopic lesions [5, 3]. Despite the high accuracies reported in these works, a visual appreciation of the output of these algorithms on video data typically shows a rather unstable output. This is mainly because these methods have difficulty handling challenges such as occlusion, deformation, motion blur, etc. These challenges are common for most video-type data and can most likely not be solved by merely increasing the number of training data for the neural network. The latest methods for video object segmentation, video action recognition and related tasks are evolving towards deep learning architectures that can incorporate temporal information for a more stable and accurate prediction [1, 4]. However, this evolution has not started in the field of endoscopy and the main reason for this is the lack of fully annotated videos which are typically needed to train such systems. Getting this kind of annotations is very costly, time-consuming and generally not feasible to obtain from a clinical expert.

We adapt a semi-supervised video object segmentation technique and use it to go from only a few manual annotations per video clip towards fully annotated clips. This effectively gives us a dataset that can be used to train this kind of temporal neural networks.

2 Data and methods

To show the concept, we have picked the task of polyp detection and have collected complete colonoscopy videos from 329 patients with a total of 605 polyps. A short video clip of ± 5 seconds (∼125 frames), containing the first apparition of each polyp is extracted and for each clip, 3-5 frames are manually delineated by a clinical expert. This results in 2.876 annotated frames showing the 605 polyps from different angles.

We use the method of [6] for semi-supervised segmentation of the entire video clips based on the manual annotations. This method uses a pretrained segmentation network and fine-tunes this on the first frame of a video to be segmented using the ground truth pixel mask for this frame. It then inferences the following frames from the video and adaptively fine-tunes the network with the most confident regions from the newly segmented frames. Here, this technique is used with three main additions: (1) we perform an additional pretraining step to adapt the network towards the domain of endoscopic video. We pretrain the model with a polyp segmentation task based on all the manual annotations that we have, (2) we do not fine-tune the network on the first frame of the video clip during inference but on all 3-5 manually annotated frames for that specific clip and (3) an additional post-processing step is performed in order to filter out any sequences that are unstable over time. The Dice score is calculated between predictions of sequential frames and this score is then used as a measure for temporal stability. Only the most stable sequences should be used for training a temporal neural network.

3 Results and conclusion

After forwarding all video clips through the network, the 2.876 manual annotations have been automatically expanded to a densely annotated dataset with 131.619 frames. A visual inspection of the results showed an overall outstanding performance of the method. Our first subsequent experiments showed that already just the increase in number of training data gives a significant performance boost when training a standard CNN for polyp segmentation. An additional merit of this annotated dataset is that we have video clips with sequentially annotated frames, which allows to train recurrent neural networks (RNNs) and other types of network architectures that require this kind of training data. In [2] we show that training RNNs is feasible for endoscopy video with the use of a sequentially annotated training set as described here.

References

1. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description (2015)
2. Eelbode, T., Demedts, I., Bisschops, R., Roelandt, P., Hassan, C., Coron, E., Bhandari, P., Neumann, H., Pech, O., Repici, A., et al.: Incorporation of temporal information in a deep neural network improves performance level for automated polyp detection and delineation. Gastrointestinal Endoscopy (2019)
3. Seguí, S., Drozdzal, M., Pascual, G., Radeva, P., Malagelada, C., Azpiroz, F., Vitrià, J.: Generic feature learning for wireless capsule endoscopy analysis. Computers in biology and medicine (2016)
4. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory (2017)
5. Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., Baldi, P.: Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology (2018)
6. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation (2017)