Semantic analysis of real endoscopies with unsupervised learned descriptors

O. León Barbed[1] LEON@UNIZAR.ES
Cristina Oriol[1]
Pablo Azagra[1]
Ana C. Murillo[1]
[1] *DIIS-i3A, University of Zaragoza, Zaragoza, Spain*

Editors: Under Review for MIDL 2022

Abstract

This work explores automatic analysis of medical procedure recordings, in particular, endoscopies. Regular medical practice recordings are noisy and challenging to process, so a quick and automatic overview of their content is essential. We show how advances in unsupervised representation learning can be applied to real medical data, obtaining rich descriptors to perform automatic semantic analysis of these recordings.

Keywords: Endoscopy, unsupervised learning, image description, scene classification.

1. Introduction

Endoscopies are a frequent medical practice producing large amounts of video. This data is tedious to process manually, and expensive to label for supervised machine learning techniques to automatically process it. Current trends tackle this problem with unsupervised machine learning techniques, producing recent advances for representation learning. We base our study on the BYOL framework Grill et al. (2020), that presents an efficient unsupervised training strategy. Our goal is to use these descriptors to facilitate semantic analysis of complete endoscopy recordings, as depicted in Fig. 1. This analysis can guide practitioners or algorithms for specific tasks, e.g., 3D reconstruction (Lamarca et al., 2020), to help processing the large amounts of data recorded. It is essential to know whether to ignore or focus on certain parts of the recordings, e.g., large intervals are useless due to poor visibility, blur, etc. Unsupervised learning of descriptors avoids bias towards pre-defined classes, and allows us to discover existing scene types. Then, an expert, with minimal supervision effort, can decide which types are of interest to be identified automatically. Our contributions are

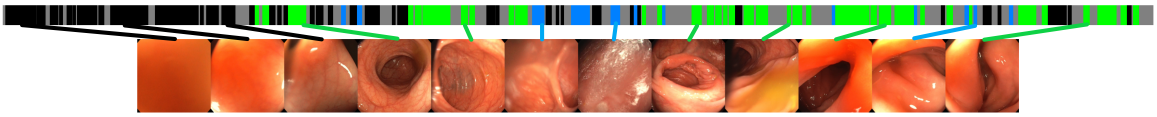


Figure 1: Top: Automatic video segmentation example (Black=*Wall*, Blue=*Water*, Green=*Good view*, Gray=*None*). Bottom: sample frames from some intervals.

an image description model[1] trained following BYOL on real medical practice data (Azagra et al., 2022), and the validation of these learned descriptors to perform recognition tasks on real endoscopy videos. We run proof of concept experiments for scene classification and semantic video segmentation.

2. Method

The BYOL **unsupervised representation learning** framework trains an online network to predict the output of a parallel target network for a different augmented view of the same source image. It uses contrastive learning but only with positive samples, which makes the training more robust and efficient. We run BYOL on a standard ResNet50 architecture (initialized with ImageNet weights) on our *train-set*. BYOL adds a projection and prediction MLPs at the end of the network that are discarded after training. The final 2048 long descriptor is the output of the ResNet50 architecture.

We build a simple **video segmentation approach** with the learned descriptors. First, we explore the type of scenes in these colonoscopies by clustering the descriptors of *train-set-2*. Visually inspecting the frames in each cluster, we observe useful semantic distinctions. In our proof of concept we select three types (classes): *Wall* represents frames where the endoscope is facing a wall, losing all visibility; *Water* are frames captured while the water pump is being used, also losing most visibility; *Good view* refers to frames where the endoscope is well positioned to examine and navigate the colon. Cluster centroids are used as a reference model to classify each frame of the video *test-set* as follows. The *classification of a single test frame* is obtained by computing the minimum distance to each class: $d^l = \min_{n=1...N_l} \left(\mathrm{L2}(x, c_n^l)\right)$, where d^l is the distance from the frame to class l, N_l is the number of centroids that correspond to class l, $\mathrm{L2}(a, b)$ is the euclidean distance between a and b, x is the descriptor of the frame and c_n^l is the n-th centroid of class l. Then we apply a filter for robustness to assign the label: if the distance to the closest class is smaller than 0.95 times the distance to the second closest class, the frame is classified as the first class. Otherwise, we add a class *none* for these uncertain cases. The final *video segmentation*, is obtained after post-processing the per-frame classification with a voting-based sliding window. The window middle frame is assigned to the most frequent class within the window.

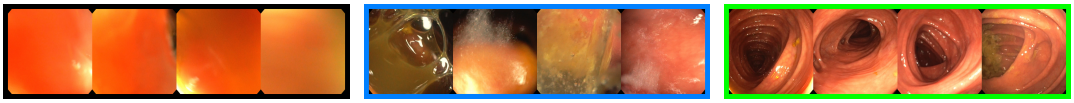
3. Experimental Results

We use the EndoMapper (EM) dataset (Azagra et al., 2022): **train-set** (379181 frames from 7 videos) and **train-set-2** (20864 frames from other 5 videos). We test on a **test-set** of 3 additional videos from EM, and the public Hyper-Kvasir (HK) (Borgli et al., 2020).

HK contains a set of *labeled-images* with frames from 23 heterogeneous semantic classes, such as polyps or cecum, that we consider as different scene types. We split the frames with a 5-fold strategy for train and test and run a MLP classifier with a hidden layer of 256. We obtain an average accuracy per class of 56.3 (random is 4.3), in comparison to 55.8 if we use a default descriptor from ResNet50 (ImageNet weights). This experiment checks that discriminative capabilities are not lost with the additional training on our data with BYOL. The performance is good even though we trained on different endoscopy data than HK.

1. https://github.com/LeonBP/VideoSegmentation

Our second experiment validates the proposed semantic video segmentation method. The clustering step is a K-Means ($k = 50$) run on **train-set-2**. Visual inspection of resulting clusters shows that a dominant class emerges for a significant number of clusters, see examples in Fig. 2. After a label is assigned to selected clusters, we apply the proposed segmentation method to the *test-set*. One example of the segmentations obtained is shown in Fig. 1. The sample interval frames point a correct classification, and we inspect that the classification results match the content of the video: frequent *wall* segments in the first part, until the endoscope reaches the deepest parts of the colon, where the doctor uses the *water* pump to clear a region. The doctor then withdraws the endoscope maintaining a *good view* to explore the parts that were not visible while entering.



Cluster 2: *wall* | Cluster 23: *water* | Cluster 19: *good view*

Figure 2: Sample images and label assigned to three clusters using the learned descriptors.

4. Conclusions

This work shows a proof of concept to semantically segment endoscopy videos, from minimal human supervision, based on a representation learned unsupervisedly. The results are consistent with the contents of the videos, which shows the promising future for this line of research. This quick overview of endoscopic videos can facilitate automatic processing of large datasets. The results and the structure of the method allow for diverse future improvements, including larger training sets, where a richer set of class labels can emerge, and more sophistication in the classification and smoothing methods.

Acknowledgments

This project has been funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 863146 and Aragon Government FSE-T45_20R.

References

Pablo Azagra et al. Endomapper dataset of complete calibrated endoscopy procedures. *arXiv preprint arXiv:2204.14240*, 2022.

Hanna Borgli et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 2020.

Jean-Bastien Grill et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020.

Jose Lamarca et al. Defslam: Tracking and mapping of deforming scenes from monocular sequences. *IEEE Trans. on robotics*, 2020.