Surgical Tool Presence Detection for Cataract Procedures

Manish Sahu, Sabrina Dill, Anirban Mukhopadyay, Stefan Zachow

Zuse Institut Berlin, Berlin, Germany

Abstract

This article outlines the submission to the CATARACTS challenge for automatic tool presence detection [1]. Our approach for this multi-label classification problem comprises labelset-based sampling, a CNN architecture and temporal smothing as described in [3], which we call ZIB-Res-TS.

I Introduction

This challenge's dataset consists of 50 videos of cataract surgeries, in which up to three different tools of 21 tools in total can be present in a frame. Each frame has a definition of 1920×1080 pixels and the frame rate is approximately 30 frames per second. The frames were annotated by two experts, where a tool is regarded as present, if it touches the eyeball. When the experts disagreed, the label 0.5 was assigned. The videos were divided into 25 training and 25 testing videos.

II Implementation Details

II.1 Pre-processing

We further divided the training videos into a training and a validation subset, using videos *train03*, *train08*, *train10* and *train21* for validation, thus ensuring that during training every tool is present. We sampled every fifth frame of the videos for training and validation, resized them to 480×270 pixels and subtracted the mean for normalizing. Due to tool usage imbalance in the training and validation videos, labelset-based sampling [3] was used to yield more balanced data. For the testing videos all frames were extracted.

II.2 Training

We employed random flipping and cropping for data augmentation and a pre-trained 50-layer residual network (ResNet-50) [2] with ImageNet weights. For fine-tuning to the tool annotation task we added a global average pooling layer followed by a fully connected output layer with 22 units, including a *no tool*-label as described in [3], on top of the ResNet-50 model. We did not consider the frames annotated with 0.5. A weighted sigmoid binary cross-entropy loss function was used. Training was performed with Keras, the Adam optimizer with a learning rate of 0.001, a batch size of 10 and for 25 epochs on a GTX1080 Ti.

II.3 Post-processing

Assuming that tool transitions in the surgical videos are smooth, linear temporal smoothing [3] was applied in order to reduce false tool detections. This approach considers a window of 29 frames (the current one and the preceding frames) with corresponding normalized, linear weights for determining the output prediction of the current frame.

III Further Notes

Our model was trained exclusively on the challenge dataset. During inference five frames per second were processed.

References

[1] CATARACTS: Challenge on Automatic Tool Annotation for cataRACT Surgery, 2017. URL `https://cataracts.grand-challenge.org/`.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[3] Manish Sahu, Anirban Mukhopadhyay, Angelika Szengel, and Stefan Zachow. Addressing multi-label imbalance problem of surgical tool detection using CNN. *International Journal of Computer Assisted Radiology and Surgery*, 12(6):1013–1020, 2017.