CRACKER - Tool identification on cataracts surgery videos with field knowledge-based filtering

Teresa Araújo*, Guilherme Aresta, Adrian Galdran and Aurélio Campilho

I. Introduction

CATARACTS, commonly described as a clouding of the eye lens, are a visual problem associated with the increase on life expectancy. In advanced stages, cataracts are treated by surgically replacing the eye lens. This surgical procedure is broadly composed of four major steps: 1) incision and drainage of the lens region, 2) destruction and aspiration of the natural lens, 3) placement of a new, artificial lens and 4) re-filling of the lens region.

Framed on the CATARACTS challenge [1], the method CRACKER (CataRACts KnowlEdge Resnet) for the prediction of surgical tools in eye surgery videos is proposed. Given the high complexity of the task, this approach is composed of an initial tool detection using a state-of-the-art multiclass convolutional neural network followed by a field knowledge-based temporal filtering.

II. Initial frame classification

The initial dataset was downsampled to 128×128, and initially distributed into a training and a validation subsets, keeping 80% of the data for training purposes. The well-known resnet34 model [3] was trained by first optimizing the cross-entropy mis-classification loss in the final layers of the network. When a low error was reached, we progressively unfreezed previous layers and continued training. The learning rate was set to $l = 0.1$ in the last layers, and progressively decreased for the previous layers. The model was initialized to weights trained on ImageNet.

The training was stopped when we observed that the validation loss was not improving anymore. The predictions for each frame on the set of test videos were built with test-time augmentation: we fed the model with four different versions of each frame, and average the resulting prediction to obtain a final probability of the presence of each tool. These predictions were then submitted to a specialized temporal filtering process, described in the next section.

III. Temporal post-processing

The main frame-wise classification procedure herein used does not have in account the temporal dimension. However, surgical procedure is expected to follow a standard set of steps, with tools appearing at different time points during different time intervals. Consequently, a post-processing that accounts for temporal information should be able to improve the CNN predictions. In order to do this, 21 1D signals are built for each video by concatenating all frame-wise prediction probabilities. These signals are then filtered to remove outliers and then processed using knowledge related to cataracts surgery.

A. Median filtering

First, median filtering is applied to each tools' signal in order to remove punctual noise. The size of the median filter window is determined based on the frame rate and knowledge regarding the human visual system. Since the average human reaction time is of 300 ms [2], and the frame rate of the videos is close to 30 fps, one can expect that the frequency of appearance/disappearance of a tool from the video to not be higher than 9 frames. In order to provide a small safety margin, the length of the window is set to 11. This filter should be able to remove isolated predictions of a given tool which most likely correspond to noise.

Considering field knowledge, there are some tools that remain for a higher period of time in contact with the eyeball than others. For these, a median filter of much larger size can be safely applied. In this work, a 101-length median filter is applied to the tools 12, 13 and 15, which correspond to irrigation/aspiration handpiece, phacoemulsifier handpiece and implant injector, respectively.

B. Inter-tool-based filtering

To further filter the signal more field knowledge regarding eye catarats surgery is applied. Some rules regarding the time points of occurrence of the different tools are established, considering inter-tool information. For instance, the irrigation/aspiration handpiece (12) and the vitrectomy handpiece (14) usually proceed the phacoemulsifier handpiece (13) because the latter is used for the destruction of the lens. Similarly, the implant injector (15) can never come before the irrigation/aspiration handpiece (12) or the vitrectomy piece (14), since the implant can only be injected into the eye once the previous damaged lens has been removed. Finally, the Rycroft cannula (4) can never come before irrigation/aspiration handpiece (12) or the vitrectomy piece (14) since this canula

T. Araújo (tfaraujo@inesctec.pt), G. Aresta and A. Campilho are with Faculdade de Engenharia da Universidade do Porto (FEUP) and INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência. A. Galdran is with INESC TEC. This work is financed by the ERDF European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme, and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia within project CMUP-ERI/TIC/0028/2014. Teresa Araújo is funded by the FCT grant contract SFRH/BD/122365/2016. Guilherme Aresta is funded by the FCT grant contract SFRH/BD/120435/2016.

is only used in the final part of the surgery to re-fill the lens region. With that in mind, the presence probability p of a tool is set to zero is set to 0 before (or after) the first occurrence of $p_{12} > 0.5$ or $p_{14} > 0.5$, whichever comes first. Please note that both (12) and (14) are very obvious tools with long in-surgery duration and consequently the identification of a correct frame should be trivial to the system, which ensures that no tool is deleted.

References

[1] https://cataracts.grand-challenge.org/

[2] https://www.humanbenchmark.com/tests/reactiontime/statistics

[3] K. He *et al.*. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/CVPR.2016.90