Towards RObust tooL Identification for cataract Surgery

G. Maršalkaitė, J. Bialopetravičius, and J. Armaitis*
Oxipit, UAB

This short writeup accompanies our second submission to CATARACTS (Challenge on Automatic Tool Annotation for cataRACT surgery). We call our algorithm TROLIS.

I. INTRODUCTION

The dataset for this challenge consists of 50 videos of cataract surgeries, split into the train (25 videos) and test (25 videos) sets. Each frame of the train dataset is labelled by two human experts, and the labels correspond to the surgical tools in use. A tool is considered to be in use as it touches the eye. 21 different tools (classes) are featured in this challenge.

Two challenges are immediately clear when it comes to the dataset. First, adjacent frames in each video are very similar. Second, the train dataset is highly imbalanced, with some tools appearing only in a single video for less than a thousand frames, whereas some other tools appear many times in each video.

II. OUR APPROACH

We split the tool categories into two lists: rare tools (6 categories: biomarker, Troutman forceps, needle holder, suture needle, Mendez ring, and Vannas scissors) and regular tools (remaining categories).

We address the first challenge by averaging each 3 frames, comparing the pixel-wise distance between them and discarding the frames which are not sufficiently different in this way. We do this only for the regular tools. We then oversample the rare tools, and undersample the empty (no tool) frames. In order to address the second challenge, we treat the rare tools in a special manner (see below).

For the regular categories we train three convolutional neural networks. In particular, we train the Resnet50 [1] on frames resized to 256x256 (nets A and B) and 512x512 (net C) pixels. In all cases we start from ImageNet weights, and account for the means of the different color channels of the images. When it comes to augmentations, we zoom in or out by up to 10%, shifting the center of the zoom either vertically and/or horizontally by up to 10%. We also flip the image horizontally, rotate it up to 10 degrees, vary the brightness, and contrast by up to 10%. All of these augmentations are performed randomly. The nets A and C are trained on all but videos numbered 4, 12, and 21, setting the aforementioned videos aside for validation (same split as team DResSys-v3). In this submission we minimize the binary crossentropy loss. Net A is trained with a learning rate of 0.1 for 900 iterations, 0.01 for 300 iterations, 0.001 for 100 iterations with SGD and a batch size of 252 (using batch accumulation). Net B is trained with the same schedule. Net C is trained with a learning rate of 0.1 for 600 iterations with the same batch size. We subsequently average the output of the three networks, apply probability clipping of 0.0005, and perform a time averaging with a window of 45 frames. We have also noticed some frames with video artifacts (tearing). We have detected these frames using a simple edge-detection algorithm. We have discarded such frames.

For the rare tools, we have made an unsupervised (classical computer vision) algorithm which detects the biomarker. Moreover, we resort to separately-trained neural networks for the rare categories. The biomarker detection algorithm works by finding black blobs (tip of the marker) and white blobs (bulk of the marker) in each frame. If these blobs are sufficiently close to each other and of a reasonable size, the algorithm outputs a positive result. We fit these parameters (the distance and the size of the blobs) on the training set video where the biomarker is present. We further assume that if the marker is used in a surgery, it is used to put at least two markings (dots on the eye). This translates to a restriction that at least 30 frames in a row should be detected as positive, and this should occur at least twice in a video in order to be counted as a biomarker instance. In order to detect the other rare tools, we finetune net A on a custom dataset. This dataset consists of 3000 frames where no rare tools are present (negative samples), 2500 frames with rare tools (positive samples). To these frames from the trainset, we add 1200 frames from the testset (negative samples). These last 1200 frames are obtained in an unsupervised manner from the testset by performing a forward pass (using net A) and taking the frames with the highest rare tool scores. We then alter net A

* info@oxipit.ai

by changing the number of its outputs from 21 to 6 (corresponding to the rare tools), and finetune it using a learning rate of 0.1, batch size 252, and 5000 iterations. Biomarker predictions of this finetuned network are discarded.

We assume that Mendez ring only appears in videos where biomarker is present, and needle holder only appears in videos with suture needle. Finally, we clip the first and last 0.5% frames of every test video.

III. RELATION TO PREVIOUS SUBMISSION

Data preparation and training procedures are different from our previous submission. We also do not use the LSTM network any more, and have added an unsupervised algorithm for biomarker detection among other changes.

IV. ADDITIONAL INFORMATION

Our system does not use any additional training data. Predictions for a given frame depend on the vicinity of the frame in question. When it comes to computational efficiency, we estimate more than 10 frames per second, depending on the hardware. The algorithm has not been tested on other datasets.

[1] K He, X Zhang, S Ren, and J Sun, "Deep residual learning for image recognition," arxiv (2015).