Artificial Intelligence Project

Semantic Segmentation of CATARACT Dataset

Submitted by: Team 8

May 6, 2020

Nikita Rana (00101032017)
Sakshi Gupta (00201032017)
Iti Shree (03401032017)
Mugdha Goel (04401032017)

Under the supervision of

Mr. Rishabh Kaushal
Assistant professor

Department of Information Technology
Indira Gandhi Delhi Technical University For Women

Contents

1 Introduction

Video signals provide a wealth of information about surgical procedures and are the main sensory cue for surgeons. Video processing and understanding can be used to empower computer assisted interventions (CAI) as well as the development of detailed post-operative analysis of the surgical intervention. Computer assisted interventions (CAI) have the potential to enhance surgeons' capabilities through better clinical information fusion, navigation and visualization. Currently, CAI systems are used mainly as tools for preoperative planning and translation of such plans into the procedure through surgical navigation. There are possibilities to develop CAI further with more advanced deformable navigation capabilities, better imaging and robotic instrumentation.

1.1 Problem statement

A fundamental building block to such capabilities of CAI is the ability to understand and segment video into semantic labels that differentiate and localize tissue types and different instruments. Data driven machine learning techniques and deep learning, in particular, have been immensely influential in recent vision advances as well as in medical image computing and analysis. Deep learning has advanced semantic segmentation techniques dramatically in recent years but is fundamentally reliant on the availability of labelled datasets used to train models. The generated dataset is conveniently separated in three different challenges with the aim to address three challenges of CAI systems: anatomical understanding, instrument identification and tracking, and understanding of interactions between surgical instrumets and anatomical landmarks.

1.2 Objective

Advancing the state-of-the-art on CAI systems- We introduce a semantic segmentation dataset built on top of the CATARACTS data. We demonstrate how this dataset can be used to train state-of-the-art deep learning frameworks for semantically segmenting unseen cataract data. We believe this can help in the development of CAI techniques based on vision.

Even though cataract surgery is less prone to complications, a small improvement and risk mitigation can have big impact. We therefore generate this dataset to foster more research on developing CAI systems for cataract surgery, which can potentially reduce risks and improve the workflow.

2 Literature Survey

2.1 Research questions

1. Are state-of-the-art models able to learn accurate anatomical representations in cataract surgery?
2. Can we achieve high segmentation accuracy?
3. What is the potential of semantic segmentation as instrument identification and tracking CAI systems?
4. What is the correlation between different instruments on cataract surgery?
5. What are the challenges of accurately differentiating different instruments? for example multiple cannulas with different surgical function look very much alike
6. how deep neural networks can perform well on image segmentation with a difficult dataset such as the one proposed here?

3 Proposed Methodology

3.1 Dataset

CaDIS: a Cataract Dataset for Image Segmentation, is a dataset for image segmentation created by Digital Surgery Ltd.. CaDIS consists of 4738 images from the 25 videos on CATARACTS' training set.

The CATARACTS challenge training set includes 25 videos that have around 500K frames in total. Because pixel-level labeling is time-consuming and the change among consecutive frames is subtle, we use ground-truth tool and phase information to select frames that have tools and are evenly distributed across different phases. As a result, we collect around 200 frames per video and 4738 frames in total.

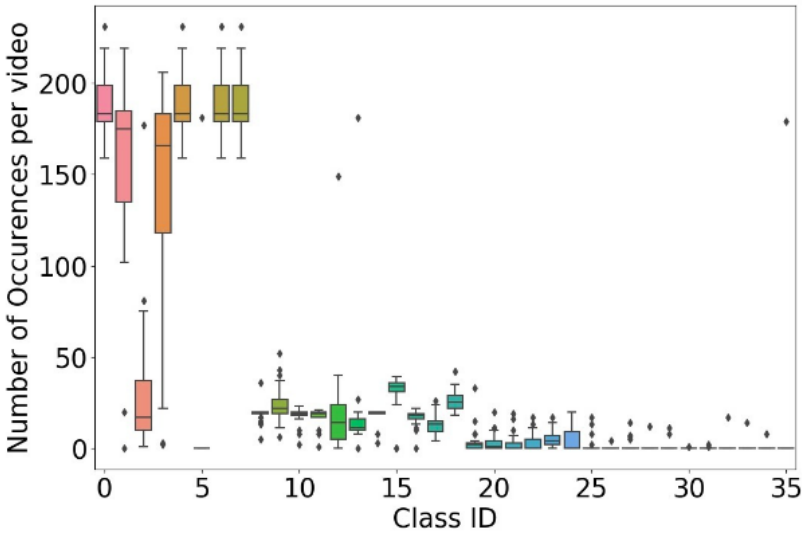
The dataset includes 36 different semantic classes: 28 surgical tool classes, 5 anatomy classes, and 3 miscellaneous classes. Surgical tool classes also include surgical tool handles: When surgical handles appear in some of the images, they were given a different class ID. The handles where given a different class ID.

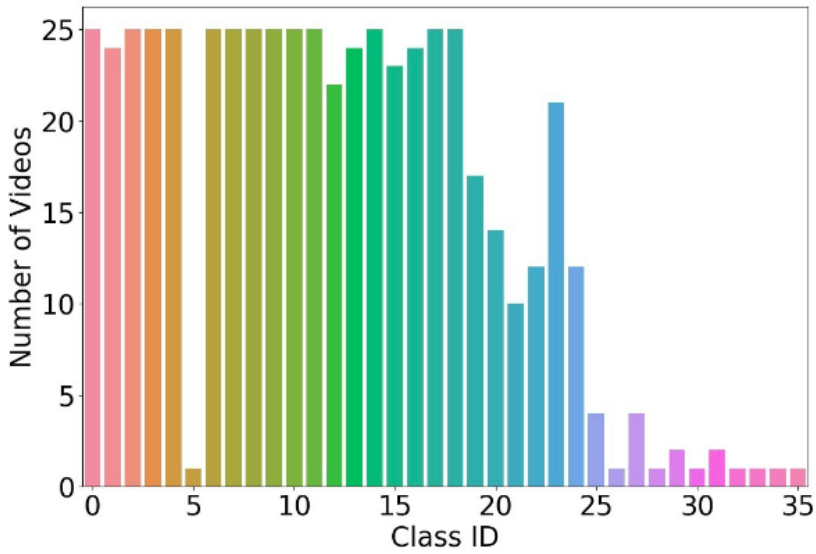
Number of Instances- 4738 images
Label Information- Label Present

Tools and Handles	Anatomy	Misc.
8. Hydro. Cannula	0. Pupil	1. Surgical Tape
9. Visco. Cannula	4. Iris	2. Hand
10,23. Cap. Cystotome	5. Eyelid	3. Eye Retractors
11,21. Rycroft Cannula	6. Skin	
12. Bonn Forceps	7. Cornea	
13,31. Primary Knife		
14,22. Phaco. Handpiece		
15,25. Lens Injector		
16,19. A/I Handpiece		
17,24. Secondary Knife		
18. Micromanipulator		
20. Cap. Forceps		
26. Water Sprayer		
27. Suture Needle		
28. Needle Holder		
29. Charleux Cannula		
30. Vannas Scissors		
32. Viter. Handpiece		
33. Mendez Ring		
34. Biomarker		
35. Marker		

3.2 Data Visualization

Boxplot(a)-provides average and scatter of the number of instances of the respective classes in one video
Barplot(b)- describes number of video in which each class appeared in at least once





Boxplot (a) provides the average and scatter of the number instances of the respective class in one video. Barplot (b) describes the number of videos each class appeared in at least once. The plots show that the data is very biased towards anatomy classes where they appear in almost all videos and having a high number of instances per video.

3.3 Details About the Organization

CaDIS: a Cataract Dataset for Image Segmentation, is a dataset created by Digital Surgery Ltd. Digital Surgery is a health tech company, based in London, UK, shaping the future of surgery through the convergence of surgical expertise and technology. The Innovation team is working on bridging the gap between Artificial intelligence and the OR. They released CaDIS to public believing a semantic dataset will encourage the computer vision community to push surgical research further.
CATARACTS: Data set was approved in April 2010 by the Information Standards Board (ISB) as an inherited information standard based on good evidence of its use a) in electronic cataract care records and b) to support national audit, benchmarking, research, and quality improvement.

3.4 Data Pre-processing

1. Data Files Combining

we use phase annotation from to split videos(e training set includes 25 videos) into 14 surgical phases. We then randomly select a maximum of 20 frames per phase such that the frames are at least three seconds apart and have a tool in it. As a result, we collect around 200 frames per video and 4738 frames in total.

2. Data Cleaning
The image are downsampled by half from 1920 1080 to 960 540.

3.5 Feature Computation

In order to evaluate the $meanIOU = \frac{(GT \cap Pred)}{(GT \cup Pred)}$ presented models, we use two metrics. The first metric is pixel accuracy that is computed as the percentage of correctly classified pixels and it is defined as: $PixelAcc = \frac{(GT \cap Pred)}{GT}$ where GT and Pred stands for ground truth and predictions respectively and indicates the intersection operation.
The second metric is the mean Intersection over Union (IoU). The mean IoU tends to penalize incorrect detection more than pixel accuracy by considering both intersection between prediction and ground truth as well as the incorrect predictions along with missed ground truth pixels. Mean IoU is defined as: $meanIOU = \frac{(GT \cap Pred)}{(GT \cup Pred)}$

4 Data Exploration

4.1 Metrics

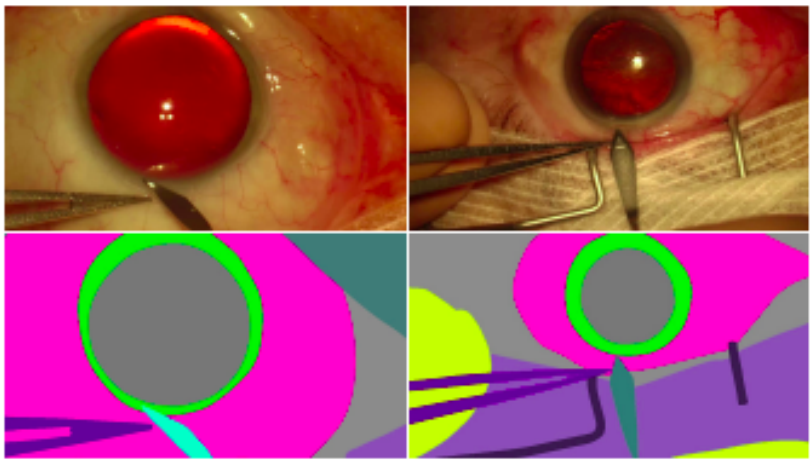
1. Anatomy Understanding
In this experiment, all instrument classes found in Table I were merged into one class leading to a total of 9 classes. This experiment focused more on the anatomy classes to help CAS applications enable anatomy evaluation during surgery and thus helping in risk avoidance and evaluation of surgical skill.

Model	Mean IOU	Pixel Acc.
PSP	26.33	45.81
DeepLab v3+	25.95	45.68

2. Instrument Identification
In this experiment, all the anatomy and misc. classes found in table I were merged into one class leading to a total of 22 classes. This experiment focused on surgical instruments during cataract surgery to pave the way towards tool usage, tool tracking and cross-tool interaction applications.

Model	Mean IOU	Pixel Acc.
PSP	22.19	49.43
DeepLab v3+	22.11	49.43

3. Surgical Understanding
The experiment is targeted towards detecting different class types together (tool and handles, anatomy and Miscellaneous). As only 0.23 of the dataset belongs to tool handle classes, tool tip and handle classes are merged where applicable. As a result, we got 29 different classes. A sample of the training data is shown in the figure.



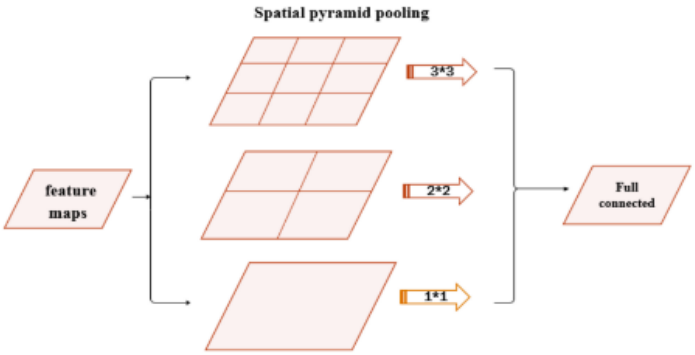
Model	Mean IOU	Pixel Acc.
PSP	34.87	45.74
DeepLab v3+	34.30	45.65

5 Algorithms

5.1 DeepLabV3

DeepLab is a semantic segmentation model designed and open-sourced by Google back. Multiple improvements have been made to the model since then, including DeepLab V2, DeepLab V3 and the latest DeepLab V3+.
The Deeplab V3 model combines several powerful concepts in computer vision deep learning—

1. **Spatial Pyramid pooling**—Spatial pyramid architectures help with information in the image at different scales i.e small objects like cats and bigger objects like cars. Spatial pyramid pooling networks generally use parallel versions of the same underlying network to train on inputs at different scales and combine the features at a later step.

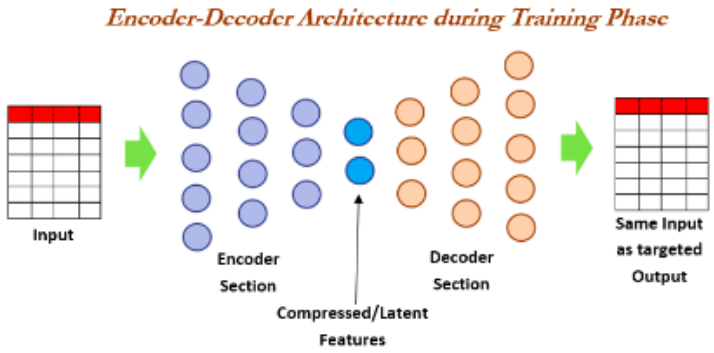


2. **Encoder-Decoder architectures**—This has become a very popular architecture for a variety of tasks in computer vision and NLP. We used the encoder to downscale the image to a feature vector that summarizes the essence of the image and then use a decoder to expand the summarized feature vector back into the dimensions of the image however the decoder would return us back an image with semantic segmentation. The DeepLab model is broadly composed of two steps:

3. **Encoding phase**: The aim of this phase is to extract essential information from the image. This is done using a pre-trained Convolutional Neural Network, now you might be wondering why a CNN? If you have previously worked with a CNN for image classification then you might know that convolutional layers look for different features in an image and pass this

information to subsequent layers, now for segmentation task what comprises the essential information, its the objects present in the image and their location and since CNN are excellent at performing classification, they can easily find out the objects present.

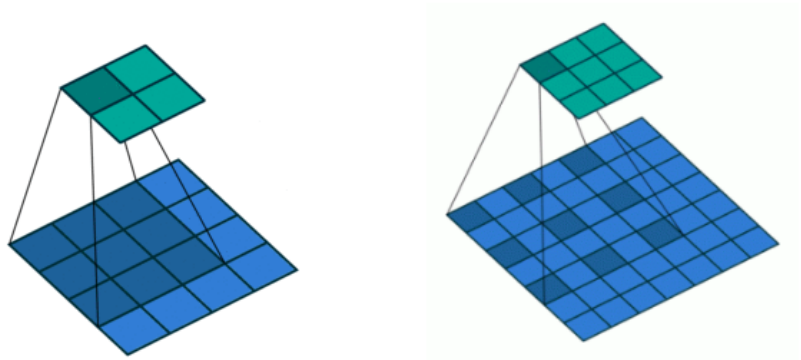
4. **Decoding phase**: The information extracted in the encoding phase is used here to reconstruct output of appropriate dimensions



5. **Atrous convolutions**—DeepLab uses atrous convolutions. Atrous convolutions require a parameter called rate which is used to explicitly control the effective field of view of the convolution. The images below shows atrous convolutions. The benefit of atrous convolutions is they can capture information from a larger effective field of view while using the same number of parameters and computational complexity

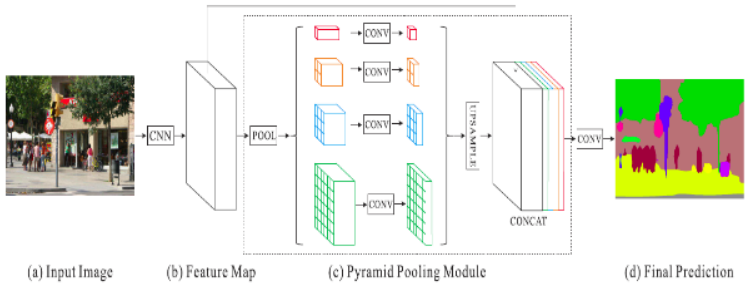
$$y[i] = \sum_{k} x[i + r \cdot k] w[k]$$

rate = 1 rate = 2



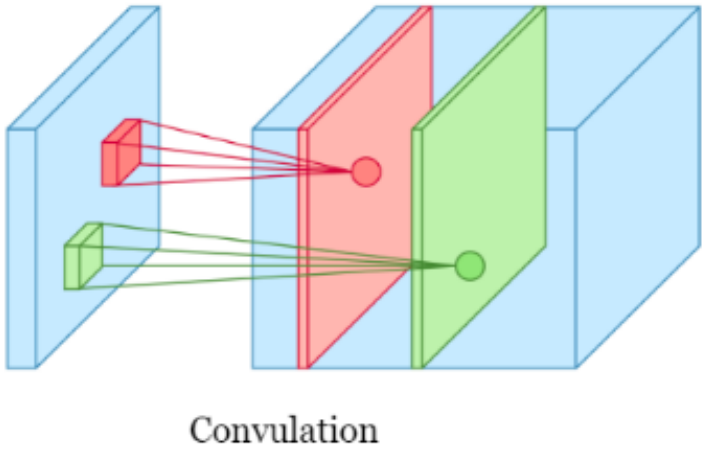
5.2 PSPNet

FCN (Fully Convolutional Network) methods have proved to have many failure cases like mismatched relationships, confusion categories and inconspicuous classes during scene parsing. Therefore, we introduce the Pyramid Scene Pooling Network (PSPNet) which proves to be an effective global contextual prior as it uses context information.
The process for PSPNet is described below in 4 steps:



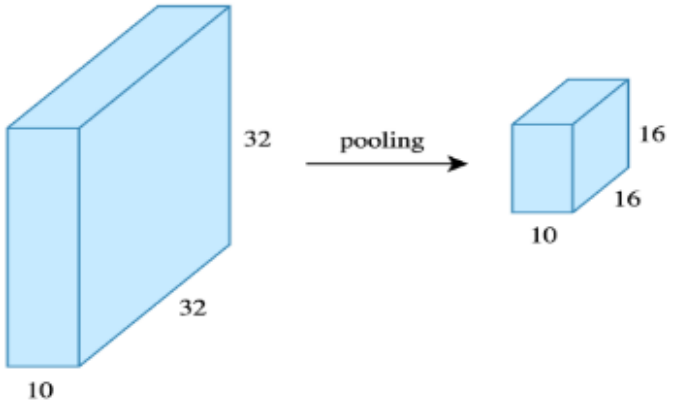
a) Input Image Input images of any shape, usually dimensions greater than (256, 256) are fed to the network.

b) Feature Map Convolution is applied on the input data using a convolution filter to produce a feature map.



c) Pyramid Pooling Module An Image contains objects of sizes ranging from small area to large area in different regions.Fully Convolution Network(FCN),U-Net and other networks constructs feature maps by

upsampling and doing segmentation at different levels for segmentation of all objects in all regions.But in PSPNet to correctly segment all size objects, feature maps are pooled average pooled at different pool size.



c.1 Sub-Region Average Pooling Sub-region average pooling is performed for each feature map.

Red: This is the coarsest level which performs global average pooling over each feature map, to generate a single bin output.

Orange: This is the second level which divides the feature map into 22 sub-regions, then performs average pooling for each sub-region.

Blue: This is the third level which divides the feature map into 33 sub-regions, then performs average pooling for each sub-region.

Green: This is the finest level which divides the feature map into 66 sub-regions, then performs pooling for each sub-region.

c.2 11 Convolution for Dimension Reduction Then 11 convolution is performed for each pooled feature map to reduce the context representation to 1/N of the original one (black) if the level size of the pyramid is N. If the number of input feature maps is 2048, then the output feature map for 4 levels in total (red, orange, blue and green) will be $(1/4)2048 = 512$, i.e. 512 number of output feature maps.

c.3 Bilinear Interpolation for Upsampling: Bilinear interpolation is performed to up-sample each low-dimension feature map to have the same size as the original feature map (black).

c.4 Concatenation for Context Aggregation All different levels of upsampled feature maps are concatenated with the original feature map (black). These feature maps are fused as global prior. That is the end of the pyramid pooling module at (c).

d) Final Prediction Finally, it is followed by a convolution layer to generate the final prediction map.

6 Bibliography

1. https://www.analyticsvidhya.com/blog/2019/02/tutorial-semantic-segmentation-google-deeplab/
2. http://hellodfan.com/2018/07/06/DeepLabv3-with-own-dataset/
3. https://developers.arcgis.com/python/guide/how-pspnet-works/
4. https://medium.com/analytics-vidhya/semantic-segmentation-in-pspnet-with-implementation-in-keras-4843d05fc025
5. http://blog.qure.ai/notes/semantic-segmentation-deep-learning-review