INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

Semantic Segmentation of CATARACT Dataset

(AI Project Report)

Submitted By
Nikita Rana (00101032017)
Sakshi Gupta (00201032017)
Itishree (03401032017)
Mugdha Goel (04401032017)

Under the supervision of
Rishabh Kaushal
Assistant Professor
Information Technology

STUDENT UNDERTAKING

Dated: 22nd May, 2020

This is to undertake that the work titled Semantic Segmentation of CATARACT in this Minor Project Report as part of 6th Semester in B.Tech. (Information Technology) during January – May, 2020 under the guidance of Mr.Rishabh Kaushal is my original work.

The report has been written by me in my own words and not copied from elsewhere. This report was submitted to plagiarism detection software on ___ (date) and percentage similarity found was ___, similarity report attached as Appendix.

Anything that appears in this report which is not my original has been duly and appropriately referred / cited / acknowledged. Any academic misconduct and dishonesty found now or in future in regard to above or any other matter pertaining to this report shall be solely and entirely my responsibility. In such a situation, I understand that a strict disciplinary action can be undertaken against me by the concerned authorities of the University now or in future and I shall abide by it.

Nikita Rana,
Sakshi Gupta,
Iti Shree,
Mugdha Goel

Date-of-Submission - 26th may 2020, New Delhi

DEPARTMENT OF INFORMATION
TECHNOLOGY
INDIRA GANDHI DELHI TECHNICAL
UNIVERSITY FOR WOMEN
KASHMERE GATE, DELHI - 110006

Dated: 22nd May, 2020

CERTIFICATE

This is to certify that the work titled Semantic Segmentation of CATARACT data set submitted by Nikita Rana, Sakshi Gupta, Iti Shree and Mugdha Goel in this project report as part of 6th Semester in B.Tech. (Information Technology) during January – May, 2020 was done under my guidance and supervision. This work is their original work to the best of my knowledge and has not been submitted anywhere else for the award of any credits / degree whatsoever. The work is satisfactory for the award of Minor Project credits.

Mr.Rishabh Kaushal
Assistant Professor
Department of Information Technology
Indira Gandhi Delhi Technical University for Women

ACKNOWLEDGEMENT

We are very grateful to have managed to complete our project 'Semantic Segmentation of CATARACT data set'. We would like to express our greatest gratitude to our professor, Mr. Rishabh Kaushal for his guidance and encouragement in completing this project. This project could not have been completed without the effort and teamwork from our group members, Nikita Rana, Sakshi Gupta, Iti Shree and Mugdha Goel.

Nikita Rana,
Sakshi Gupta,
Iti Shree,
Mugdha Goel

Semantic Segmentation of CATARACT Dataset

Nikita Rana, Sakshi Gupta, Itishree, Mugdha Goel

May 2020

Abstract

Semantic Segmentation of CATARACT data set would enable CAI (Computer Aided Intervention) capabilities in cataract surgery. The model would semantically segment the retinal scans of cataract eye into 36 different classes of surgical tools, anatomical components of human eye and miscellaneous which would help in pre-operative and operative planning of the procedure.

We have used an already created data set CATARACTS which consists of 4378 images from 25 videos of Phacoemulsification i.e. the cataract surgery, to implement our model for semantic segmentation.

We have used two unsupervised and two supervised learning algorithms to segment the data set images. The unsupervised algorithms include Watershed algorithm and K-Means clustering and the supervised algorithms include DeepLabV3 and PSPNet. As a result of applying these algorithms to create a semantic segmentation model, we get segmented images from input cataract eye images for the different algorithms.

Contents

1 Introduction

Image segmentation is one of the most crucial steps of computer vision, including examples such as autonomous driving or extracting critical information from medical images.
A large amount of information about surgical procedures can be obtained from video signals. These signals are of great help to the surgeons as they provide main sensory data. Processing, understanding and analysis of these video signals can be used to improve computer assisted interventions (CAI) and help in the development of detailed analysis for surgical understanding.
Computer assisted interventions have a great potential to improve surgeon's capability through better medical information, navigation and visualisation. CAI systems are being used as a tool for preoperative planning and surgically navigating these plans into procedures.

1.1 Problem Statement

CAI capabilities include the ability to understand and then segment videos into different semantic labels. These labels help to differentiate between different instruments and identify tissue types. For medical image computing and analysis purposes, data driven machine learning techniques and deep learning are being intensively used. Deep learning uses labelled data sets to train models and importantly, also has advanced semantic segmentation techniques.
Here, we aim to address three different challenges of the CAI system including anatomical understanding, instrument identification and understanding of interaction between surgical instruments.

1.2 Objectives

- To compare and contrast the segmentation performed by unsupervised learning algorithms vs supervised learning algorithms.
- To advance the research in the field of computer assisted interventions (CAI).
  We aim to demonstrate how CaDIS dataset can be used to train deep learning frameworks to semantically segment the given cataract data which in turn may lead to reducing potential risks and improving the surgical workflow.

1.3 Motivation

Increasing number of cataract cases and complications arriving from across the globe have led to the need of developing new methodologies for cataract surgery with minimal risk and high precision. The high computation power and flourishing field of computer vision has made the possibility of such CAI features very real today. In this project we have studied the effectiveness of various unsupervised and supervised learning algorithms for image segmentation.

2 Literature survey

2.1 Research Questions

1. Are the state of art CAI models able to learn anatomical representations seen in cataract surgery accurately?
2. Can we achieve high segmentation accuracy?
3. What is the potential of precise segmentation of various instruments identification?
4. What are the challenges for successfully differentiating between different instruments? For example, many cannulas(a surgical instrument used in cataract surgery) have different surgical functions but look alike.
5. What is the correlation between different instruments of cataract surgery?
6. How well can a deep neural network perform on the image segmentation of this difficult data set?

3 Related Work

There have been previous work in the field of cataract detection, such as analysis and study of cataract detection techniques, automatic cataract detection using Deep Convolutional Neural Network and studies on models which helps in segmentation such as DeepLab V3+, PSPNet, UNet etc.

Dr. Maria Grammatikopoulou's [6] paper also extends the data set use to check the accuracy of various models.

3.1 Comparison of Proposed Approaches

The approaches are based on state-of-the-art models which provides baseline for semantic segmentation. We used PSPNet, DeepLab V3+, K-Means and Watershed to segment the images and calculate the heuristics. DeepLab V3+ is an extension of DeepLab that uses modified Xception and uses it after combining it with atrous convolution. With different dilation rates it achieves better contextual predictions with the same image resolution.

3.2 Research Gaps

In this paper we have presented the effect that the CaDIS data set has on the semantic segmentation of the CATARACT data set, unlike any previous work.

4 Proposed Methodology

4.1 Dataset Description

CaDIS - Cataract Data set for Image Segmentation.
This data set is created by Digital Surgery Ltd. and is used for image segmentation purposes. CaDIS has a cataracts' training set of 25 videos which containing 4738 images.
The videos have approximately 500K frames in total. Ground truth tool and phase information is used to shortlist frames which have tools and are evenly distributed across various phases. This is done due to time consuming nature of pixel level labeling and subtle changes in consecutive frames. Therefore, we finally have 200 frames per video and a total of 4739 frames.
There are 36 different semantic classes, including : 28 surgical tool classes (it also includes surgical tool handles; the handles are given different class ID), 5 anatomy classes and, 3 miscellaneous classes.

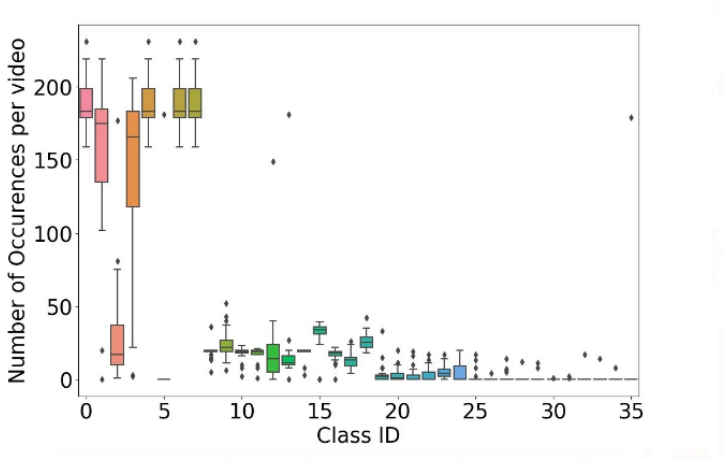
Number of Instances- 4738 images
Label Information- Label Present

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		

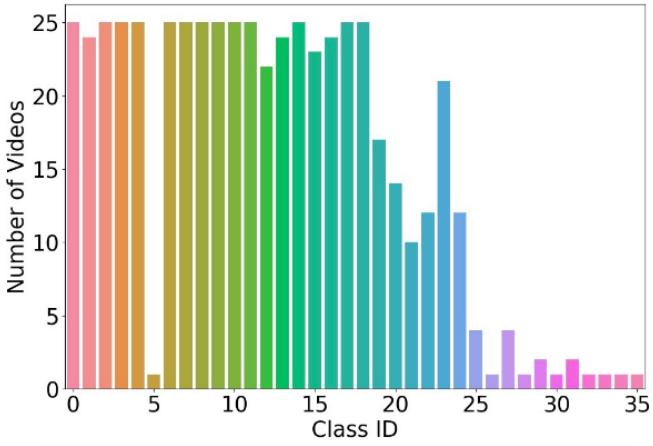
Table I

4.2 Data Visualization

We can see from the plots that the data is biased towards the anatomical class tags. The reason for this could be their high number of instances per video.



we can observe the average and scattered number of instances from the classes in video one.



we can observe the number of videos where each classes appeared at least once.

4.3 Details About the Organization

We took the data set from CaDIS: a Cataract Data set for Image Segmentation, which is a data set created by Digital Surgery Ltd. The release of CaDIS data set was made public with the belief that a semantic data set will encourage the computer vision community to work on various research topics. The Data set got approved in April, 2010 by the Information Standards Board (ISB).

4.4 Data Pre-processing

1. Data Files Combining

In this project we are performing segmentation on the data set. We had approximately 200 frames per video and 4738 frames in total.

2. Data Cleaning
We used down sampled images which were halved from 1920 * 1080 to 960 * 540.

4.5 Feature Computation

Pixel accuracy: It implies the percentage of correctly classified pixels in the segmented image when compared with ground truth image and is defined as:
$PixelAcc = \frac{(GT \cap Pred)}{GT}$
GT is the ground truth and Pred stands for predictions.
Python code to measure pixel accuracy
pix_acc = pixel_accuracy(pred_image, label_image)

5 Algorithms

5.1 Unupervised Learning

5.1.1 Watershed

Watershed is a transformation on gray scale images. The aim of watershed is to segment parts of an image where two regions to be segmented are closer to each other, that is they may share boundaries.It works on the principles of edge detection and flooding.
We perform these steps.

1. **Conversion to Gray Scale**: Input image is first converted to gray scale since watershed works on gray scale images.

2. **Thresholding**: We perform automatic thresholding of the image using Otsu's Binarization. This returns image divided into foreground and background.

3. **Noise Removal**: We remove all the noise in the gradient image. If this step is not done then it may lead to segmentation due to noise. This may lead to over segmentation.

4. **Marker Controlled Image Segmentation**: Marker is a connected component of a image. We use markers to modify the gradient image. We have two kind of markers, internal markers for objects and external markers for boundary. Markers are place inside objects and external markers are associated with the background of the image.

Step-wise Working

1. Read the original image *I*.
2. Morphological reconstruction of the *I*.
3. To detect the minimum, compute the complement of image obtained by the morphological reconstruction, the result image noted *Ic*.
4. For markers of the original image, subtract from the original image *I*, the image *Ic*: *dif* = *I* - *Ic*
5. Extended and imposed minimum, we obtained the *markers*
6. Compute the watershed transform of the markers
7. Show the watershed segmented image.

5.1.2 K-Means Clustering

K-means is an unsupervised clustering algorithm which groups similar data points into K clusters or classes.
The process starts with selecting K centroids and calculating the distance of each data point from the centroids. The data point is then added to the cluster whose centroid is nearest to that point. Repeating this process a certain

number of times results in K clusters having similar data points in each. K-Means algorithm aims at clustering the input image into k classes.

Step-wise Working

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1. Randomly select '*c*' cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is the minimum of all the cluster centers..
4. Recalculate the new cluster center using: $$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

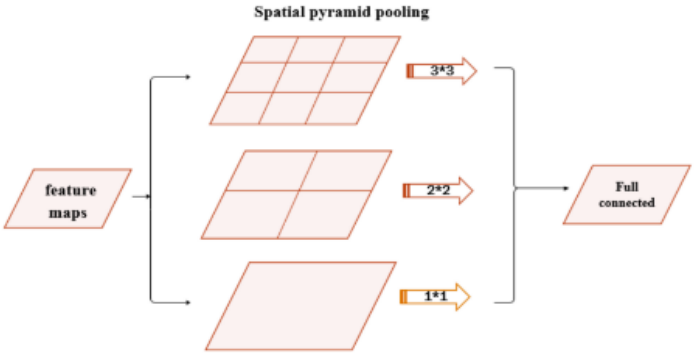
   where, 'c_i' represents the number of data points in i^{th} cluster.
5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3.

5.2 Supervised Learning

5.2.1 DeepLabV3

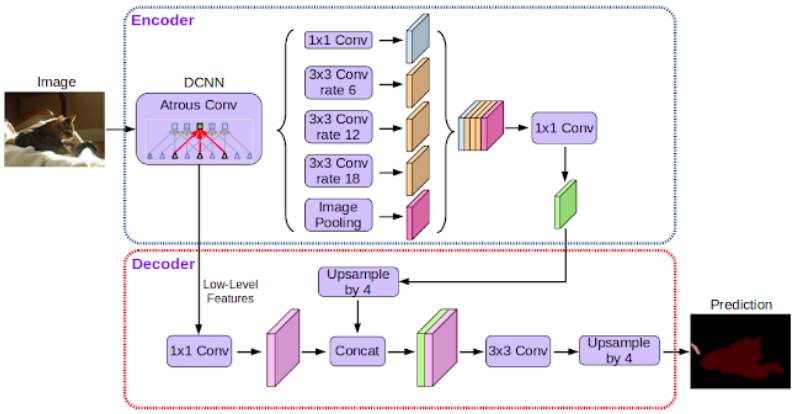
DeepLab is a semantic segmentation model designed and open-sourced by Google. In this project, we have used DeepLabV3 to perform segmentation on the CADIS data set. The Deeplab V3 model combines many powerful concepts of computer vision and deep learning, which are.

1. **Spatial Pyramid pooling**: Spatial pyramid architectures highlight the information within the image at different scales like the small class objects of surgical instruments and bigger class objects of anatomical classes. Spatial pyramid pooling allows the input image to of any sizes. This allows arbitrary aspect ratios and arbitrary scales. The input images can be re-scaled to any scale size and used with the required scale. Parallel versions of the same underlying deep network are used by the pooling network to train on inputs from images re-scaled to different sizes and then the features are combined.



2. **Encoder-Decoder architecture**: The encoder phase down scales the image to a feature vector that summarizes the crux of the image. The decoder phase expands the summarized feature vector back into the dimensions of the image. Furthermore, the decoder would return us back an image complete with the required semantic segmentation with appropriate classes.

i. **Encoding phase**: This phase is aimed at extracting essential and crucial information from the image which will lead to correct segmentation of the image in the decoding phase. This is performed on convolution neural network, convolutional layers look for different features in an image and pass this information to the subsequent layers.



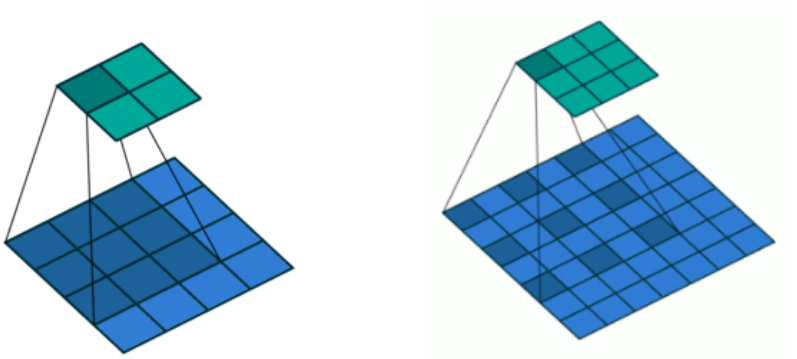
ii. **Decoding phase**: The information extracted from the encoding part is employed here to reconstruct the output of needed dimensions.

3. **Atrous convolutions**: DeepLab uses atrous convolution also known as dilated convolutions, to explicitly control the resolution at which the responses of features are computed in a deep convolutional neural network. Atrous convolutions need a parameter referred to as rate that is employed to specifically manage the effective field of vision for the convolution. Atrous

convolutions are better than traditional convolutions as Atrous convolutions capture data from a bigger effective field of vision while the number of parameters and computational complexity used stays the same.

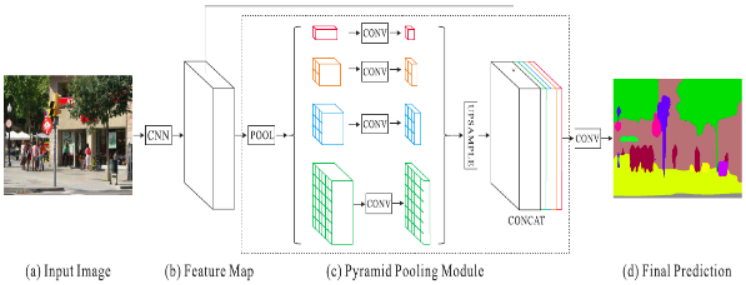
$$y[i] = \sum_{k} x[i + r \cdot k]w[k]$$

rate = 1 rate = 2



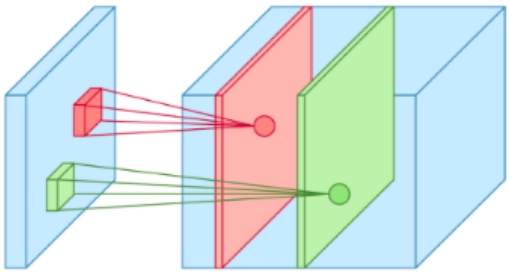
5.2.2 PSPNet

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 context information helps remove confusion in similar looking classes and avoids mismatching of classes.
The process for PSPNet is described below in 4 steps:



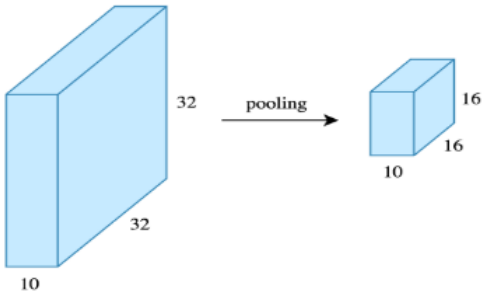
a) Input Image: Images of any shape, usually dimensions greater than (256, 256) are fed to the network as input.

b) Feature Map: A feature map is produced by applying a convolution filter on the input. Convolution is mathematical operation used to multiply to two arrays of different sizes but same dimensions.



Convulation

c) Pyramid Pooling Module: There are small-area and large-area objects in different regions of an image. Therefore, to correctly segment all the different sized objects PSPNet uses average pooling with different pool sizes on the feature maps whereas FCN, U-Net and other networks use up-sampling to generate feature maps and segments at different levels to segment different sized objects.



c.1 Sub-Region Average Pooling: Each feature map is divides into sub-regions and sub-region average pooling is applied.

Red: First level is the coarsest level which performs global average pooling over each feature map and generates an single bin output i.e 1x1.

Orange: The second level divides the feature map into 2x2 sub-regions and performs average pooling for each sub-region.

Blue: Then feature map is divided into 3x3 sub-regions and each sub-region is average pooled.

Green: This is the finest level which divides the feature map into 6x6 sub-regions and then performs pooling for each sub-region.

c.2 1X1 Convolution for Dimension Reduction: Then 1x1 convolution is performed for each pooled feature map to reduce its 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 Bi-linear Interpolation for Up-sampling: Each low-dimension feature map is up-sampled by performing bi-linear interpolation so that they are of the same size as that of the original feature map (black).

c.4 Concatenation for Context Aggregation: The up-sampled feature maps of different levels and the original feature map (black) are concatenated together to form a global prior to use as context information for segmentation. This concludes the pyramid pooling model.

d) Final Prediction: Finally, a convolution layer is applied on the global prior to generate the final prediction map.

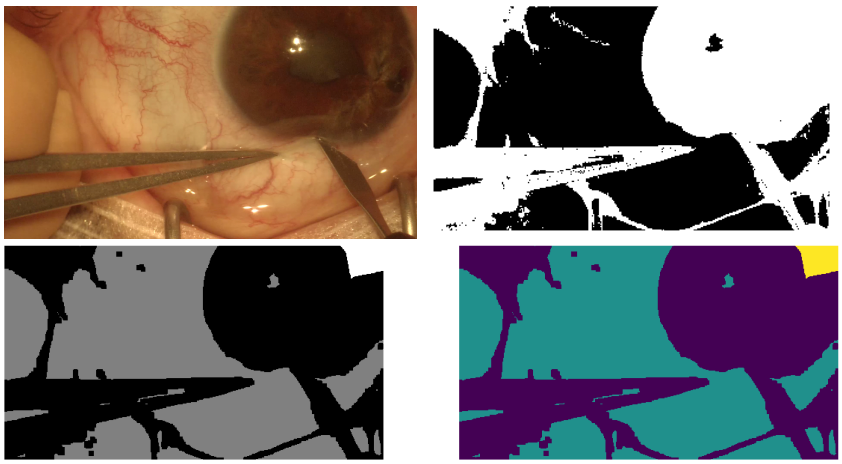
6 Experiment Setup & Results

6.1 Unsupervised Learning

For unsupervised learning algorithms, we have given the input images from the CATARACT data set and then performed image segmentation on the image.

6.1.1 Watershed

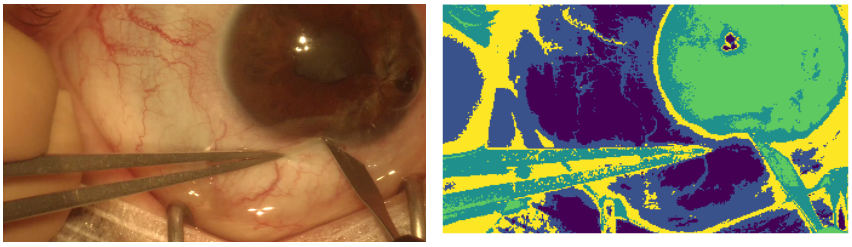
In watershed, we observed that the algorithm is able to segment the image correctly into the foreground interest objects, for example pupil, surgical instruments etc, and background.

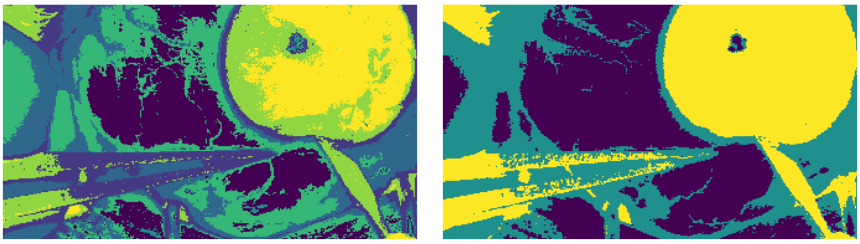


Watershed: Input image and Segmented images.

6.1.2 K-means Clustering

In k-means clustering, we observed that increasing the number of classes leads to over-segmentation in the images. K = 3 gives the most accurate results.





K-Means Clustering: Input image and Segmented images for k = 5, k = 7 k = 3 respectively.

6.2 Supervised Learning

For supervised learning, we have trained our model on the CaDIS dataset

6.2.1 Metrics

1. Anatomy Understanding
In the first experiment of the classification, all the instrument classes in the Table I are merged into one class. This leads to 9 classes in total. It helps in identifying if the segment of image is one of the anatomical class or any surgical instrument class. This experiment helps evaluate the anatomy of the eye during the surgery and therefore helps in avoiding risk and evaluating skills.

Model	Pixel Accuracy
DeepLab v3+	83.7

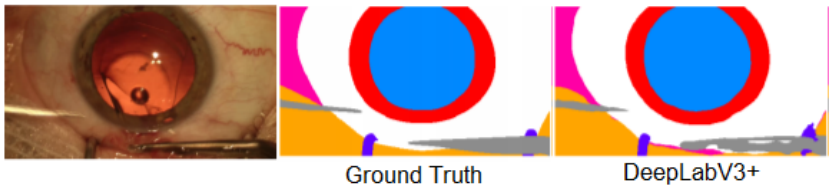


Fig. : Example frame with ground truth segmentation and model predictions for Experiment I. (Colormap: Pupil, Iris, Cornea, Skin, Surgical tape, Eye retractors and Instrument)

2. Instrument Identification
In the second experiment of the classification, all the anatomy and miscellaneous classes in Table I are merged into one class. This leads to 22 classes in total. It helps in identifying if the segment of image belongs to one of the surgical instruments or the merged class of anatomical, or miscellaneous classes. This experiment helps in tool usage, tool tracking and cross-tool interaction during the cataract surgery.

Model	Pixel Accuracy
DeepLab v3+	85.3

Fig. : Example frame with ground truth segmentation and model predictions for Experiment II. (Colormap: Pupil, Iris, Cornea, Skin, Surgical tape, Eye retractors, Cannula, Capsulorhexis cystotome, Tissue forceps, Capsulorhexis forceps, Secondary knife, Lens injector, Micromanipulator and I/A handpiece)

3. Surgical Understanding

In the third experiment of classification we aim to detect different class types together (tool and handles, anatomy and Miscellaneous). Since the tool handle classes, tool tip and handle classes comprise only 0.23% of the data set therefore, these classes are merged where applicable. This leads to 29 classes in total.

Model	Pixel Accuracy
DeepLab v3+	86.1

Fig. : Example frame with ground truth segmentation and model predictions for Experiment III. (Colormap: Pupil, Iris, Cornea, Skin, Surgical tape, Eye retractors, Viscoelastic cannula, Rycroft cannula, Capsulorhexis cystotome, Capsulorhexis cystotome handle, Lens injector, Primary knife, Bonn forceps, Capsulorhexis forceps, Micromanipulator, I/A handpiece and I/A handpiece handle)

7 Conclusion

In this work we have shown difference and contrast between the performance of different unsupervised and supervised learning algorithms for semantically segmenting images from CATARACT data set using CaDIS data set for training for supervised learning algorithms.
We have conclude that unsupervised learning algorithms give quite good result but the differentiation among classes is not quite accurate, whereas the supervised learning algorithms perform quite well with proper class identification and image segmentation.

8 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

6. https://arxiv.org/pdf/1906.11586.pdf

7. https://sthalles.github.io/deep_segmentation_network/

9 Appendix: Similarity Report

In this soft copy, attach snapshot of first page of similarity report which clearly indicates percentage plagiarism.