ISSN: 2834-7706

Research Article

Journal of Robotics and Automation Research

Improving Cataract Surgery Procedure using Machine Learning and Thick Data Analysis

Chandrashekhar Singh*, Jinan Fiaidhi and Sabah Mohammed

Department of Computer Science, Lakehead University, Canada

***Corresponding Author**
Chandrashekhar Singh, Department of Computer Science, Lakehead University, Canada.

Submitted: 2024, May 12; **Accepted**: 2024, Jun 28: **Published**: 2024, Jul 05

Citation: Singh, C., Fiaidhi, J., Mohammed, S. (2024). Improving Cataract Surgery Procedure using Machine Learning and Thick Data Analysis. *J Robot Auto Res, 5*(2), 01-08.

Abstract

Cataract surgery is one of the most frequent and safe Surgical operations are done globally, with approximately 16 million surgeries conducted each year. The entire operation is carried out under microscopical supervision. Even though ophthalmic surgeries are similar in some ways to endoscopic surgeries, the way they are set up is very different. Endoscopic surgery operations were shown on a big screen so that a trainee surgeon could see them. Cataract surgery, on the other hand, was done under a microscope so that only the operating surgeon and one more trainee could see them through additional oculars. Since surgery video is recorded for future reference, the trainee surgeon watches the full video again for learning purposes. My proposed framework could be helpful for trainee surgeons to better understand the cataract surgery workflow. The framework is made up of three assistive parts: figuring out how serious cataract surgery is; if surgery is needed, what phases are needed to be done to perform surgery; and what are the problems that could happen during the surgery. In this framework, three training models has been used with different datasets to answer all these questions. The training models include models that help to learn technical skills as well as thick data heuristics to provide non-technical training skills.

Keywords: Deep Learning, LSTM, Artificial Intelligent, CNN, Thick Data

1. Introduction

The Centers for Disease Control and Prevention data says that one in seven Canadians is predicted to experience the onset of at least one of glaucoma, retinal disorders, cataracts, or macular degeneration at some point in their lives. Given that the majority of vision loss and eye problems in Canada are preventable, these are frightening statistics. In reality, 75% of the nation's prevalent eye issues can be successfully addressed or avoided. A change in lifestyle, early detection, and treatment are frequently the keys to good eye health. Regular eye exams from an ophthalmologist are essential because so many common eye disorders and major visual issues have no symptoms when they first appear [1].

1.1. Role of AI in Cataract Surgery Training

For a long time, ophthalmology training was based on Halsted's method. In this type of training, the trainee should: spend a lot of time caring for patients under the direct supervision of a qualified surgeon; learn the science behind the disease that needs surgical treatment; and develop the skills to do operations that get more complication as they go on. Also, only people who had completed a predetermined number of procedures were considered able to accomplish surgery successful. But this method is time taking and trainee spends lots of time and energy to learn procedure because learning based only on the number of procedures done and direct practice with the patient which has some limits and risks. One of these problems is that the level of skill gained isn't always the same because there are different ways to learn, and one of the risks is that a patient might be treated by a surgeon who doesn't know what they're doing [2].

AI can be used to improve training for cataract surgery by identifying the different parts of the surgery on videos taken during surgery [3]. Videos of cataract surgery are often available to teachers and students, but they are not very useful for training right now. AI can be used to make tools that can easily break up videos of cataract surgery into its different parts so that automated skill testing and feedback can be done afterward [4]. Expertise in cataract surgery is important for public health. The cataract surgery videos are available at large scale. Surgi- cal videos vary greatly in terms of image quality, objects in the field, and movement artifacts [5]. With the help of thick data analytics and using deep learning, and Convolution neural network, we are creating a framework for cataract surgical workflow.

1.2. Surgeon Actions During Cataract Surgery

Prior to the cataract surgery the surgeon makes a painless ultrasound examination to examine the size and shape of your eye a week or two before surgery. This aids in choosing the proper lens implant type IOL. These lenses sharpen your eyesight by concentrating light on the retina. It becomes a permanent component of your eye and needs no maintenance. Throughout the process the eye will first be dilated with eye drops by your surgeon. Local anesthetics will be used to numb the region, and could also be given sedatives to make more comfortable [6]. Surgeon follows the following sequences of phases to perform cataract surgery: Incision, Viscous Agent Injection, Rhexis, Hydro dissection, Phacoemulsification, Irrigation and Aspiration, Capsule Polishing, Viscous Agent Removal and Tonifying, Antibiotics. These are called phases of cataract surgery.

2. Related Research

2.1. Convolutional Neural with Multi-Image Fusion for Surgical Tool Identification During Cataract Surgery

Another study recognizing tools from video surgery the authors in using a multi-image fusion technique to enhance the ability to recognize tools in cataract videos, which serves as the first stage in surgical workfow analysis [7]. In this work, rather of using a single video frame, each video is represented by a series of related frames. Authors used convolutional neural network (CNN) which contains few CNN layer. A pooling layer comes before each convolutional layer. The dropout layer is once again placed after the last two fully connected layers, which are designed to forecast tool presence. A group of 16 consecutive video frames are sent into the CNN. The optical fow between each successive image in a series is calculated, and the results of this calculation are processed by the first few layers of the CNN. The activation map for one image is fused to the activation maps of the subsequent successive images in order to combine information from one image to the next in a sequence. The activation map of the last image in the series is then fused once again with the activation map that resulted from the previous two images. The CNN is first trained on each individual video frame before being retrained on image sequences to further examine how the model's performance varies for various input data. The authors demonstrated that using an image sequence is superior to the traditional method of having a CNN analyze each individual video frame for the purpose of enhancing the tool detection performance of the CNN.

2.2. Surgical Tool Detection Using Attention-Guided CNN

In this research work Authors Used Faster RCNN to build a modulated anchoring network for laparoscopic video surgical equipment detection [8]. The three components of a modulated anchoring network are the modulated feature module, the anchor box location prediction, and its shape prediction. The purpose of the anchor position branch is to produce a probability map, and the map indicates the potential location of any item or instrument's center. The anchor shape prediction branch attempts to determine the shape of the instrument at the position of the discovered anchor box. To further compare the features with the shape information of accessible tools in the operation, a modified feature module combines the shape information of the instrument or item provided by the shape prediction branch into a feature map. A relation module that tries to calculate the relative relationship of various instruments in every situation of the videos is incorporated in the network. Authors used ResNet-101 to combined with a Feature Pyramid Network is used to create the fundamental feature detection network [9, 10]. Each surgical instrument in a video frame receives bond- ing box labeling from the modulated anchoring network. By creating heat maps for each surgical instrument, the authors investigated the movement of the devices that had been observed further. By analyzing tool use patterns using heat maps and the chronology of how long each instrument is used throughout a video, it is possible to assess the operational efficiency in surgical videos. The author demonstrated that Faster RCNN-based networks outperformed other current methods in terms of tool identification accuracy.

3. Using Thick Data Analytics to Implement the Framework

Thick data is made up of qualitative information, like observations, feelings, and reactions, that shows how peo- ple feel in their everyday lives. Thick data tries to find out about people's feelings, stories, and models of the world they live in. Thick data allows us to gain a deeper understanding of a dataset. While big data is a lot of complicated, unstructured information, it is large in size, and to extract meaning and support information, fur- ther preprocessing is needed for unstructured and semi- structured data sources, including text, audio, and video. To perform any deep learning task, we need a lot of data, which is either very expensive or not readily available. Thick data uses smaller samples of data to find patterns, whereas big data uses a lot of data to find patterns at a large scale in deep learning. We designed our surgical workflow framework using thick data analytics. The use of transfer learning enables one to avoid the requirement for a large amount of new data becasue in transfer learning, a model that has already been trained on a task with plenty of labeled data [11]. Transfer learning is one way to reduce the size of the datasets needed for training a deep learning model. Understanding the depth of insight in our dataset, such as surgeon experience, and the emotions of trainees, helped us design our framework perfectly. The surgeon experience heuristic assisted in the creation of a good dataset, and the model result was better at predicting the correct sequence of the surgery phase.

4. Developing Framework for Cataract Surgery Workflow

In the field of ophthalmic surgery, tiny tools are used to perform cataract surgery while looking through a mi- croscope. Retina laser treatments, refractive surgery, and cataract surgery are all examples (lens replacement). Even though ophthalmic surgeries are similar in some ways to endoscopic surgeries, the way they are set up is very different. For endoscopic surgeries, an endoscope sends a video to a large screen in the operating room, which the surgeon uses to control the procedure. But the endoscope can also be a source of information for a group of stu- dents who are observing the operation to learn. This isn't possible for ophthalmic surgeries

because the operating surgeons use a microscope that only lets one more trainee watch the surgery (through an additional ocular). This makes it hard to teach and train young surgeons, which is especially frustrating because ophthalmic surgery is one of the most difficult types of surgery and requires special operation techniques and psychomotor skills that need to be trained intensively. The surgery is microscopic and recorded for scientific, educational, and documentary purposes [12]. Using this video, we can create frameworks for surgical workflow analysis of ophthalmic surgery. The frameworks will accelerate learning process for students or surgeon. This will help understating the recognition of cataract surgery workflow such as given a scene from ophthalmic surgery, what's happening and what tools being used.

First, we need to check if the patient has a severe cataract or not. If they have severe cataracts, they def- initely need surgery. If surgery has been initiated, then how many phases of cataract surgery? What tools are being used in each phase, and what complexities could be encountered while performing the surgery? Using this framework, we can provide trainees to understand the cataract, surgery phases, and complexity. All these an- swers will be given by this framework. There could be many complexities, but during surgery, detection of the iris and pupil is difficult. because the tools used in surgery make the iris and pupil unstable.

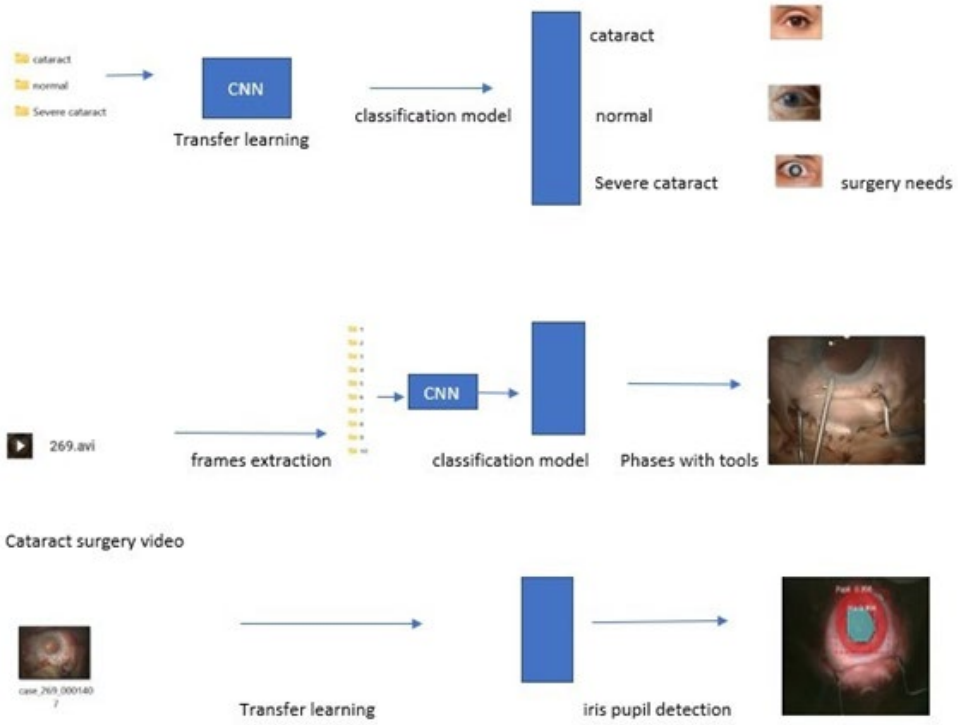

Figure 1: Framework with Three Different Models

In this framework, there are three steps, and in each step, different datasets and models have been used. Mainly, we have used thick data analytic using transfer learning to perform each task such as severe cataract detection, phase detections, and surgery complexity.

4.1. Identify Actions in a Video to Analyze It and Predict the Right Sequence of Surgery Phase

Action recognition is the process of recognizing different actions from a series of two-dimensional frames in a single video clip, where the action may be performed in full or in part. Action recognition is mainly a transformation from image classification to the classification of a series of images in a video [13]. Frames are also images. Sequences of frames turns into small clip of videos. When we work with video data set, we create frames(images) to perform any deep learning task. We designed python algorithm to extract 500 frames of each phases from Cataract 101 dataset to create a training dataset for phase extractions task eg Fig 1.We are feeding this dataset to the model to train it so that it can recommend the correct sequence of phases of cataract surgery.

4.2. Convolution Neural Network

CNNs are powerful image processing algorithms that use deep learning to do tasks like generative and descriptive. This algorithm is used to process and recognize image and video activities.A CNN is a multilayer perceptron and is better than a conventional artificial neural network, which needs less computational power. A CNN consists of three layers: an input layer which takes input like images, an output layer which provides ouput of images, and a hidden layer with multiple convolutional layers, pooling layers, fully connected layers, and normalization layers.

4.3. Long Short-Term Memory (LSTM) and Transformer Model

LSTM is advanced version of RNN which helps to remember

past data in memory. It is based on encoder- decoder method and capabilities to retain long mem- ory. LSTM solved one issue in RNN is vanishsing gradient. LSTM has ability of learning long term dependencies. It was designed to overcome short-term dependency prob- lems. It is capable of retaining informations for long time. But there are downside of it. LSTM takes long time to train as result it requires more memory to train. It has tendency to overfit and dropout is much difficult to imple- ment. LSTM is also considered slow as it does too much computation.In addition to that It is recursive nature and cannot be trained parallel. A new architecture is intro- duced to solve such a complex task called Transformer.The first time the transformer was talked about was in the paper "Attention is All You Need", which was about translating languages. The architecture of the transformer is very complicated. But the most important part is the idea it is attention based models [14]. A transformer is type of neural network that discovers context and subsequently meaning by tracing relationships in sequential data, such as the words in the phrase.Transformers are one of the newest and most powerful types of models that have been made so far.

Step 1. Image Dataset for Cataract Detection

This dataset has three types of label, no cataract, cataract and sever cataract. The dataset has been anno- tate manually. The dataset has more than 50 images of each category.

1) Training and model compilation: The dataset is divided into train and test dataset.Used data augmen- taion process to make our training more robust. Since we have small dataset, so we have used Nasnetlarge pretrained CNN model for feature extractions.Then pre- pared Sequential model with 8 layers.Compiled model with

100 epochs, Adam optimizer and loss function categori-cal crossentropy. The result comes out with 70% accuracy rate.

Step 2. Video Dataset for Phases Detection

The dataset used in this work is taken from ITEC (In- stitute of Information Technology). The dataset Catarct- 101 made up of videos of 101 cataract surgeries done by four different surgeons at the Department for Ophthal- mology and Optometry at Klinikum Klagenfurt Austria's largest public hospital. These videos were collected and annotated by a senior ophthalmic surgeon with different phases of catarct surgery [15]. All of the videos have a PAL resolution of 720x540 pixels and are encoded as MP4 files using H.264/AVC with profile High as the video codec (25 frames per second, about 1.25 MBit/s bitrate) [12]. There are three CSV files that describe the Cataract 101 videos dataset: annotations. csv, phases.csv, and videos.csv. The Phase.csv file contains only name of phases from 1 to 10. and video.csv file describes Video ID, Number of video frames, surgeion experience (1 is low experince and 2 is high experience). The surgery video's frames and phases are described in the annotation csv. In that file, there are three columns named Video ID, Frame No, and Phase. Using this dataset, we are creating training dataset which will help to predict right sequence of cataract surgery.

4.4. Creating A Dataset for The Phase Detection Model by Utilizing Surgical Video Performed by A Less and More Ex-Perienced Surgeon

We have created handy python algorithm to extract frames of corresponding phases. We used annotation.csv and video.csv field to create a our training dataset. By utilizing this algorithm, we created two types of datasets from surgery videos completed by less experienced and more experienced surgeons. we used video ID 269 (accom- plished by less experience surgeon) and 350 (completed by high experience surgeon) Let's see step by step how we created an algorithm for frame extractions. We used only one video (Video ID 269 or 350) to create data for further processing. For example, We used cataract surgery video ID 269 which was completed by less experience doctors. Let's say we have video number 269; the start frame is 68 for phase 1, and the end of frame number for phase 1 is 1043. So here we have extracted 500 frames between 68 and 1043 for phase 1 and stored them in a folder "named phase 1". I fol- lowed the same steps for the other phases. We have extracted 500 frames per second and stored them in a folder using create Dir method and creating a name for the correspond- ing phase eg phase 1,2,3 and so on. Fig 1. described the second steps. Our training dataset is ready to use our model.

```
import os
import cv2
import pandas as pd
import numpy as np
def c r e a te D i r (d i r e c t o r y):

i f not os. path. e x i s t s (d i r e c t o r y): os. makedirs (d i r e c
t o r y)
df ← pd. Readcsv (r′annotations.csv′, sep ←′;)
df ← df [df [′V ideoID′] ← 269]      using video 269
FrameLists ← df [′FrameNo′]. to list () cap ← cv2.V ideoCapture
(r′case 269.mp4′) phase ← df [′Phase′].to list ()
width ← int (cap.get (cv2.CAP PROP FRAM E W IDT H)) eight
← int (cap.get (cv2.CAP PROP FRAM E HEIGHT) size ←
(width, height)
f ourcc ← cv2.V ideoW riter f ourcc (′XV ID′) framecount ← 0
start ← 0
end ← 0
idx ← 0
while cap.isOpened() do
ret, frame ← cap.read()
if framecount in FrameLists then
idx ← FrameLists.index(framecount) start ← framecount
createDir(os.path.join(r′cataract101/′, str(phase[id])) out ← cv2.V
ideoW riter(os.path.join((r′cataract101/′
, str(phase[idx]),
"video.avi"), f ourcc, 20.0, size)
if idx ≤ len(F rameLists) − 1 then
end ← FrameLists[idx + 1]
else
end ← start + 500
end if
```

if start≠0 & end≠0 & framecount ≤start+500 **then**
out.write(f rame)
else
framecount+ = 1
end if
if idx ←len(FrameLists)-1 & framecount ←end **then**
break **end if** cap.release() cap.release()
cv2.destroyAllW indows()
end if end while

4.5. CNN-LSTM and Transformer Implementation for Phase Detections

The CNN-LSTM model can be created by first adding CNN Layers, then LSTM Layers with Dense Layer on the output. In the first experiment, LSTM was used for phase predictions. We used InceptionResNetV2 for fea- ture extractions with "imagenet" weights and used the String Lookup layer to turn the class labels into number.

While the Transformer model is also another type of neural network that contains an encoder/decoder archi- tecture like LSTM or other RNN. In the experiment, we used a different feature extractor, which is DenseNet121. The input size and weights of the CNN model were the same as we used for InceptionResNetV2.

4.6. Training and Model Compilation.

The whole dataset was divided into test data, test labels, train data and train labels. The size of the feature frames in the train set is (10,500 ,1536) and frame mask feature in train set is (10,500). First, we initialized frame feature input with param- eters (MAX SEQ LENGTH, NUM FEARURES) and mask as input with parameter (MAX SEQ LENGTH, type bool). The input of the first layer of LSTM was frame features and masks (labels). In all the layers, the activation function was used and also added dropout in the layers with a value of 0.1. The final layer has softmax as an activation function. The model was compiled with Adam optimizer. The cost function we have used is sparse categorical crossentropy. This loss function is used for multi-class classification models where the output label is given an integer value, like our dataset.

In case of Transformer model, we have again used the String Lookup layer to turn the class labels into numbers. First, the self-attention layers that make up a Transformer don't care about sequence so, we need to use positional embedding. Since videos are made up of a series of frames OR frames are always in order taken from videos, our Transformer model needs to take this into account. We used an positional embedding layer to store the positions of the frames in a video. Then, we added these positional embeddings to the CNN feature maps that have already been made. In simple words, we can say that positional embedding layers take care of position of an input vector, which is very important if we have data in a sequence. Other than Positional Embedding layers on top, we added GlobalMaxpooling1D layer which takes the max vector over the step dimension and is followed by a dropout layer at a rate of 0.1 and a final layer added with softmax as an activation layer. We used Adam as the optimizer and the same loss function, which is sparse categorical crossentropy. We used both types of dataset and found that experience surgeon dataset model was better than less experience sugeron dataset.

Step 3. Dataset for Iris and Pupil Detections

We used another dataset from ITEC for detection of complexity during cataract surgery, which keeps
82 frames from 35 videos of cataract surgery, and this dataset is also annotated with areas of iris and pupil [12]. The dataset was recorded at Klinikum Klagenfurt (Aus- tria). The resolution is 540X720pixels with frame rate of 25 fps. Since this dataset was taken during cataract surgeries, images have different surgical tools and depict the eyes in different states, such as with artificial lenses, without lenses, and tools used. This dataset is also anno- tated and it is in COCO Format. The dataset is already annoted and divied into test, train and validation. The COCO dataset is in JSON format. The format contains five sections of information for the entire dataset such as Info, Licenses, Categories, Images, Annotations. We only need Images, Annotations, and Categories fields for our task. Image contains information about the dataset; Anno- tations contain annotation information such as bounding boxes and Categories explains the labeling of images [16]. To proceed with our experiment, we built custom iris- pupil dataset for iris to train our model.

4.7. Build the Custom Iris and Pupil Dataset Like Coco Dataset and Trained Using Mask R-CNN

Before training, we have built an iris pupil cus- tom dataset. For this work, we created a classclass which is derived from the DATASET class. This class helped to process the iris pupil JSON data. DATASET class allows for the simultaneous loading of many object from JSON data. This class is really use- ful and provide different helpful methods to work on. There are different methods in this class that have been used such as load mask, load datset, add class and add images. load dataset method helped to iter- ate through all object of dataset such as image ob- ject, annotation object and categories in Json file and using add class and add image, a custom dataset was created. Another important class that we used is load mask class. This method generates masks for every object in the image, such as the mask over the iris or pupil. It will return class ids, one mask per instance, and a one-dimensional array of class ids for the mask instance. Later The dataset was divided into validation, test, and training datasets. We used the Mask-R-CNN respository, which is MatterPort Mask R-CNN, to work on this dataset We loaded pre- trained weights for Mask R-CNN from COCO data for better results [17]. We run through 250 epochs at a learning rate of 0.0001. Once model training was completed, we saved this model in a folder and later used this trained model for prediction.

5. Results and Discussions

5.1. CNN-LSTM and Transformer Result Analysis Results Analysis with Less Experience and High Experience Surgeon

CNN-LSTM	Transformer	
Phase	**Accuraye(%)**	**Accuraye(%)**
1	18.37	25.67
2	8.21	10.35
3	6.33	4.30
4	5.97	0.41
5	6.66	2.45
6	7.90	12.25
7	4.31	7.61
8	7.61	17.74
9	13.93	0.88
10	20.71	18.33

Table 1: Predicted Result Metrics of CNN-LSTM and Transformer with less Experience Surgeon

We used data of phase 1 from test dataset. From Figure 4, we can see that predicted result metrics are provided by the trained model (less experience dataset). In simple words, it describes the possibilities of a predicted video clip that belongs to phase 1. It described that there was an 18.37% chance that it belonged to phase 1, while a 20.71% chance that the test video clip belonged to phase 10. On the other hand, we can see that the Transformer model predicted a better result than CNN-LSTM and that it described the correct result. It described that 25.6%7 chances, which was higher than other phases values, that the video clip was from phase 1. But again, the model was confused about Phase 1 and Phase 10. This model has said that this video clip could be phase 10 because the model provided a higher chance value, which is 18.13% after Phase 1 while Phase 1 and Phase 10 are verydifferent steps which also cannot be repeated.

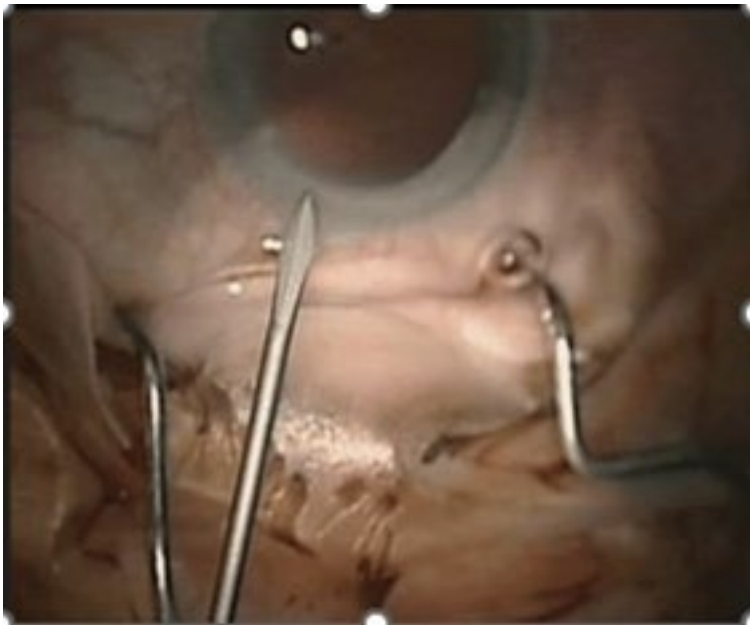

Figure 2: Predicted Result

Less Ex Type of dataset CNN-LSTM	Transformer Accuraye(%)	Accuraye(%)
Perience dataset	10	20
High Experience dataset	20	40

Table 2: Accuracy Rate of CNN-LSTM And Transformer with Two Different Dataset

From table II, we can see that, high accuracy rate of Transformer model as as we trained with dataset of experienced surgeon.

5.2. Iris Pupil Detection Results

For prediction we have used 90% detection min confi- dence. This helps us to detect bounding boxes and classes. The confidence score displays the likelihood and degree of certainty with which the classifier believes that the box contains an item of interest. The confidence score should be zero if there is nothing in that box. In our case, we have given 90 percent, which means that a model could detect objects with bounding box very less restrictively. There predicted results are in Fig 2.

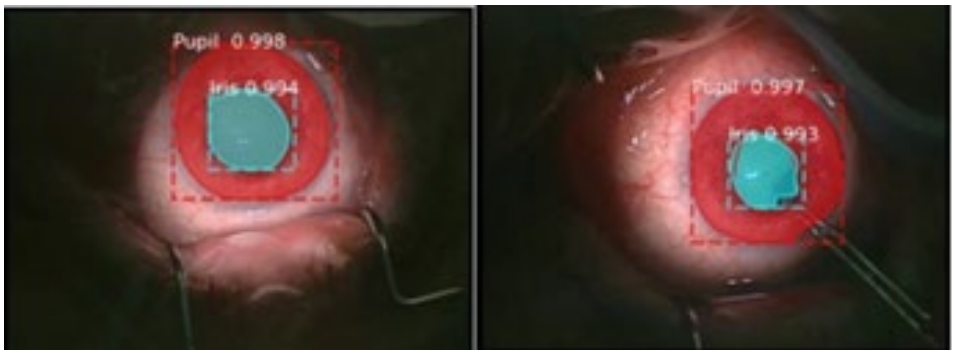

Figure 3: Iris Pupil Detection Result with Confidence Code More Than 90 Percent

The mAP metrics were calculated for Mask RNN to understand how this object detection model performed over the coco dataset. Based on the mAP metrics, the Mask RNN model has performed well. For example, in Figure, the model processed the first images taken from the validation folder with only a ground truth vect is 2, indicating that the images have two objects, iris and pupil, and a predicted vect value is 2, indicating that the model detected two objects (iris and pupil) correctly and created a bounding box over them

```
Processing 1 images
image                    shape: (512, 512, 3)         min:    0.00000  max:  255.00000  uint8
molded_images            shape: (1, 512, 512, 3)      min: -123.70000  max:  151.10000  float64
image_metas              shape: (1, 15)               min:    0.00000  max:  512.00000  int64
anchors                  shape: (1, 65472, 4)         min:   -0.70849  max:    1.58325  float32
the actual length of the ground truth vect is :  2
the actual length of the predicted vect is :  2
Average precision of this image :  1.0
The actual mean average precision for the whole images 1.0
Processing 1 images
image                    shape: (512, 512, 3)         min:    0.00000  max:  255.00000  uint8
molded_images            shape: (1, 512, 512, 3)      min: -123.70000  max:  151.10000  float64
image_metas              shape: (1, 15)               min:    0.00000  max:  512.00000  int64
anchors                  shape: (1, 65472, 4)         min:   -0.70849  max:    1.58325  float32
the actual length of the ground truth vect is :  4
the actual length of the predicted vect is :  4
Average precision of this image :  1.0
The actual mean average precision for the whole images 1.0
Processing 1 images
image                    shape: (512, 512, 3)         min:    0.00000  max:  255.00000  uint8
molded_images            shape: (1, 512, 512, 3)      min: -123.70000  max:  151.10000  float64
image_metas              shape: (1, 15)               min:    0.00000  max:  512.00000  int64
anchors                  shape: (1, 65472, 4)         min:   -0.70849  max:    1.58325  float32
the actual length of the ground truth vect is :  6
the actual length of the predicted vect is :  6
Average precision of this image :  0.3333333432674408
The actual mean average precision for the whole images 0.777777781089147
```

Figure 4: Mean Average Precision of Mask R-CNN

6. Conclusions

We offer a multitasking learning process for cataract surgery from detection to surgery. Utilizing three different dataset types, such as image datasets, video datasets, and coco-like datasets of Cataract, we were able to complete three linked tasks in the same domain. The framework will assist students and trainees expedite their training for cataract surgery. The major goal of the research was to improve the coherence of cataract instruction. A skilled surgeon can complete the treatment in 10 to 15 minutes. Surgeon residents do not get enough time to observe the process. Only the surgeon performing the surgery and a second person with an additional ocular can see the entire process because it is being done under a micro- scope. It is not similar to endoscopic surgery or other surgeries in any way. To improve training and make it simpler to understand, the major goal of this study is to mitigate issue found by new training and make their learning smooth. On the

other side, this framework is helpful for both catract patients and trainee surgeons. A person with cataracts must take a number of steps, including consult an eye surgeon, getting an ultrasound, and waiting for the surgeon's turn. All of these procedures are quite expensive and lengthy, and in many developing nations, thousands of people do get such a facility where patients can go check their eye and find cataract severity. This framework (Step 1) makes it simple to demonstrate whether or not they have a contract.

A. Limitation

The accuracy rate of CNN-LSTM is only 10% and Transformer model is only 20%. Both models were not able to perform very well on the data. We added more frames, epochs and layers in both model but did not get satisfactory result. Fig show the loss graphs of CNN-LSTM and Transformer. However, we found good results in cataract surgery severity and iris pupil detection. The reason behind having that dataset labeled properly is for iris pupil detection and cataract surgery severity identification. The model was unable to learn properly in the cases of LSTM and Transformer.

B. Future Work

The accuracy rate of the LSTM and transformer must be improved. Datasets play a crucial role in any machine- or deep learning task. In this case, a well-organized dataset for cataract surgery would be helpful for phase extractions and tool identification to improve the accuracy rate. Other than during cataract surgery, some phases are done twice. This could happen because the surgeon does not have enough experience. There could be risk and complexity during cataract surgery, such as bleeding, IOL instability, retinal detachment, and so on. A trainee surgeon should be aware of these types of problems. A new model or frame- work should be designed to address these complexities.

References

1. Common Eye Problems - Glaucoma, Cataracts, AMD More — Seema Eye Care Center (2022).
2. *Cataract surgery training around the world - Eye- Wiki.*
3. Lindegger, D. J., Wawrzynski, J., & Saleh, G. M. (2022). Evolution and applications of artificial intelligence to cataract surgery. *Ophthalmology Science, 2*(3), 100164.
4. Kawka, M., Gall, T. M., Fang, C., Liu, R., & Jiao, L. R. (2022). Intraoperative video analysis and machine learning models will change the future of surgical training. *Intelligent Surgery*, 1, 13-15.
5. Yu, F., Croso, G. S., Kim, T. S., Song, Z., Parker, F., Hager, G. D., ... & Sikder, S. (2019). Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. *JAMA network open, 2*(4), e191860-e191860.
6. *Cataract surgery – Mayo Clinic.*
7. Al Hajj, H., Lamard, M., Charrière, K., Cochener, B., & Quellec, G. (2017, July). Surgical tool detection in cataract surgery videos through multi-image fusion inside a convolutional neural network. In *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 2002-2005). IEEE.
8. Shi, P., Zhao, Z., Hu, S., & Chang, F. (2020). Real-time surgical tool detection in minimally invasive surgery based on attention-guided convolutional neural network. *IEEE Access, 8*, 228853-228862.
9. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).
10. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *In Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).
11. N, Arya. (2022). *What is Transfer Learning?*
12. Schoeffmann, K., Taschwer, M., Sarny, S., Münzer, B., Primus, M. J., & Putzgruber, D. (2018, June). Cataract-101: video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM multimedia systems conference* (pp. 421-425).
13. Tanin, U. H. (2022). *Deep video analysis methods for surgical skills assessment in cataract surgery* (Doctoral dissertation, Carleton University).
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30.*
15. Schoeffmann, K., Taschwer, M., Sarny, S., Münzer, B., Primus, M. J., & Putzgruber, D. (2018, June). Cataract-101: video dataset of 101 cataract surgeries. In Proceedings *of the 9th ACM multimedia systems conference* (pp. 421-425).
16. COCO *format*.
17. Mask, R. (2019). CNN for object detection and instance segmentation on Keras and TensorFlow. *GitHub repository.*