Deep Learning Based Multi-Label Classification for Surgical Tool Presence Detection in Laparoscopic Videos

by

ASHWIN RAJU

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2017

To my parents and my siblings.

ACKNOWLEDGEMENTS

I would like to thank my Professor who supervised my Thesis, Dr. Junzhou Huang who mentored me regularly and made me to learn and do research in a more efficient way. I would also like to thank Dr. Heng Huang and Dr. Jia Rao for serving as committee members.

It would like to express my gratitude to Mr. Sheng Wang, who was my mentor and without him it would be highly difficult to finish the thesis on time. I would like to add Mr. Zheng Xu, Mr. Viswa and other friends in my lab for encouraging and motivating to do thesis.

I would like to thank my family and relatives in India who where supporting me constantly. At last, I would thank God for giving me such sweet people for spreading positive vibrations around me.

June 26, 2017

ABSTRACT

Deep Learning Based Multi-Label Classification for Surgical Tool Presence Detection in Laparoscopic Videos

Ashwin Raju, M.S.

The University of Texas at Arlington, 2017

Supervising Professor: Dr. Junzhou Huang

Laparoscopic surgery, Modern surgery, where the surgery is performed far away from the patient by inserting small incisions on the patient's body and the surgery is performed with a help of a video recorder and through which the doctor performs the surgery. The computer assisted intervention are increasing exponentially and the need for accurate and reliable intervention is very important because of the domain which is very critical. Efforts have made to develop a system that is both fast and accurate approach but it is still an active area of research due its importance. Some applications which involve identifying the location of surgical tool at the given frame, identifying what tools are present in the given frame and many more applications. With the advance of deep learning models, the computer Assisted intervention are getting its reward and many papers have been published in this domain recently.

In this thesis, a Deep learning based multi-label classification method for identifying surgical tools in a given frame was developed and it was able to beat other methods that participated in the competition. The pipeline consists of Video to image frame conversion, Model training with real-time data augmentation, ensemble meth-

ods for combining the models. The model mainly consists of Convolutional neural network with many layers. The key concept for performing a best state of the art method was to combine the two state of the art method and evaluate the test set. We use Inception architecture and the standard feed-forward architecture for performing the prediction. This method was able to beat other results and was able to get the first place in MICCAI challenge. The results was evaluated by MICCAI conference and the data was provided by them as well.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Laparoscopic surgery

Laparoscopic surgery [1] is a modern surgery that is being used in most of the surgical operations where a small incisions are made on the patient's body and with the help of a video recorder the surgery is performed. The main advantage of Laparoscopic surgery are the fast recovery, less hospital stay and much lesser scars. There are two types of Laparoscopic surgery 1) A telescopic rod lens system, where the surgery is assisted with a video camera. 2) Digital laparoscopic surgery where the charged-couple device is placed at the end of the laparoscope. The most common Laparoscopic surgery is the telescopic rod lens system and in this type of surgery, the camera transmits the image of the organ to the video camera and the doctor performs the operation with the help of the video camera or in other words the video camera is the eye for the doctor. Various surgical instruments are insisted through the small incisions and each surgical instrument is assisted by a video camera. Laparoscopic surgery are famous for liver, pancreas, bile duct and with the help of computer assisted intervention the surgery has been explored on various organs which are much complicated. We will be talking about computer assisted intervention in this surgical domain in much detail in the future sections. Even though, there are many advantages on using this system there are few disadvantages such as risk on using small incisions and operating with the help of a video camera because the doctors has to rely on some other instrument rather than his/her eye.

[1] http://www.healthline.com/health/laparoscopy

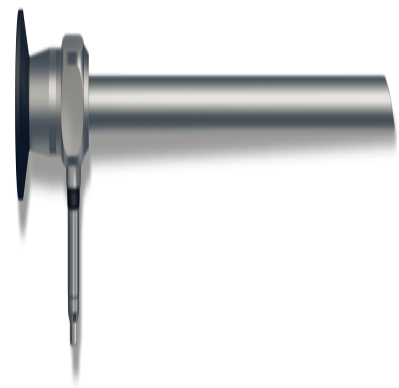


Figure 1.1: Laparascope

The computer Assisted Robotic surgery is the future of Laparoscopic surgery, where the Robots are well trained to perform surgery on the patient's body. The main advantage of this type of surgery is the Accuracy and the capacity to perform many surgeries as it does not require doctors to interfere. In order to make this happen we need the robots to be extremely trained or in others words the the error the robots make should be close to zero and at the same time the robots should be in the position to decide when an uncertainty occurs. There have been a lot of simulators that is currently being trained and this is an active area of research for the Machine learning community to work and improve the existing models. We shall discuss about the computer assisted intervention in much detail on the next sections.

1.2 Computer Assisted Intervention

CAI, Computer Assisted Interventions are an active research practice where the medical interventions are supported with the help of computer based tools and methodologies. Some examples include medical robots, surgical instruments navigation, user interface development. There have been many publications that promote the advance of Computer Assisted intervention mainly, MICCAI (Medical imaging and Computer Assisted Intervention) which organizes several competitions on handling Computer Assisted Intervention problems. We would see much detail about the competition in the upcoming sections. The general paradigm of CAI is first to gather information about the patient based on the preliminary analysis made on the patient. At the second stage, A plan of action is made and evaluated with the database that has previous or similar records. At the third stage, The action is executed . This step can be both manual or with the help of robots. Later, The results are evaluated to make sure the surgery is properly executed as planned.The future of CAI according to many researches is to make the plan more efficient, accurate and at the same time fully automatic.

1.2.1 Types of Laparoscopic Surgery

1. **Telescopic rod lens system**: Telescopic rod lens system where the surgery is done by inserting a video camera and the organs are captured by the video camera. The doctors use the surgical tools to operate and with the help of a camera there is no need for open surgery. The telescopic rod lens system is the most common Laparoscopic surgery and it makes the surgery more efficient and accurate. The recovery stage after surgery is very fast. The need for computer assisted intervention in this type of CAI is in demand because the computer can guide the doctor and can make the surgery easy. For example, the computer can give the intersection between the organ and the surgical tool which makes the doctor to find how sensitive the specific tool is respective to the organ and another example would be to find the location of the surgical tool in the image. For this kind of study, we need an advance Machine learning model which is very accurate. We will discuss about the machine learning model in future chapters.

2. **Charged-coupled device**: The charged coupled device was first introduced in 1982, which is made of silicon chip and it is attached to the end of the surgical tool. This device has better camera quality and is able to cover any angle. They have pixels, which is a silicon chip covered in image sensors and it converts incoming light energy from a visual scene into a digital signal that can be stored, processed or transmitted with greater efficiency and reliability than its analog equivalent.
3. **Future of CAI**: The future of CAI happens to be fully automated which involves the signals transferring from the Charged couple device to the system and with the trained model which operates the affected portion and decides when there is a critical situation. There will not be any intervention by the doctor.

1.2.2 Challenges in CAI

1. **General Challenges**: Even though CAI has greatly been facilitated with advance technologies, smarter and efficient devices which can visualize and give reliable data, it is still under active area of research. There had been papers and conferences conducted specially on this particular domain because of the demand and its necessity. General challenges in CAI would be accuracy and reliability. From the accuracy perspective we would want the model to be a good balance of recall and precision or in other words we should not have errors. From the reliability perspective, we want the CAI to handle uncertainty as well as take smart decisions during critical situations where the doctors themselves are not in the position to answer.
2. **Machine learning challenges in CAI**: Machine learning plays an important role in this intervention. Some of the very useful applications are identifying

the surgical tools, converting low resolution images into high resolution images, removing variances in the images, localizing surgical tools and many more. With the advance of deep learning models the CAI becomes more state of the art and it sometimes defeats the human accuracy. we will look into deep learning and its state of the art methods in the future sections.

3. **Future of ML in CAI**: The future of Machine learning would be unsupervised approach and using deep learning. The main use of unsupervised approach is to avoid feeding annotations to the machine as it is labor intensive. The unsupervised are becoming famous in deep learning and the researches expect that it would become state of the art in the near future.

1.2.3 Surgical Tools

1. **General introduction**: Surgical Tools is used inside the incisions to operate the organs. The Surgical tools are specially designed for Laparoscopic surgery as it should be very flexible and easy to handle for the doctors. Some of the surgical tools are scissors, specimen bag, clippers. These tools are made of stainless by adding nickel and chromium in measured qualities. These tools last for a long time and they are small and compact. The surgical tools and CAI comes in handy because without surgical tools the CAI does not have a good role and without CAI the doctors might find it difficult to perform surgery. We will see the need for CAI with respect to surgical tools in the next sections.

2. **Classification**: Classifying surgical tools in a given frame is very important for the robots as they can predict the next tool that is to be inserted for a new given surgery. The robots are trained such that when a new test video is passed it identifies what tools are to be inserted in that particular frame. Classification

of surgical tools is the main focus of this thesis and we would be looking into different methods that is the state of the art methods to classify surgical tools.

3. **Localization**: Localizing surgical tools is another important role for Machine learning with respect to surgical tools. The localizing is useful for robots to identify the position and angle of the specific tool in a given video frame. The position of the tools has to be accurate because the organs can be sensitive and any small error might affect the organ on the whole. Localizing is the extended approach of classification, Machine learning models can do both localization and classification at the same time and with the same model.
4. **Need for Machine learning**: Classification or Localization can be done by manual feature extraction/ computer vision but the accurate prediction of these tools with respect to varied image quality becomes difficult for computer vision, techniques to identify and localize the tools. Even standard Machine learning methods fails to achieve this goal and we will move to deep learning based methods which happens to be the state of the art.

1.3 Problems & Challenges in surgical Tools detection

Surgical tool detection is an active area of research and there has been a lot of papers published in this domain. Even though there are lot of active research, there are still on going problems and challenges in this domain. The main problem would be the localization of tool where the images have a lot of variance. The angle of camera can be different at different places and the position of instrument can be varied. The model must be able to predict and identify the location in a most accurate way. The model has to provide a very good accuracy because even a small loss in accuracy would result in loss of life. The other challenges include the unsupervised learning during uncertainty. The model must be in a position to analyze a different situation and react smartly which is similar to expert's decision. We will be solving classification task in this challenge which involves classifying the different surgical tools in a given image.

1.3.1 Current status in CAI

The current status in Computer Assisted Intervention is deep learning but even then it has been fully explored. There has been a relatively simple approaches that has been explored and and has made a promising baseline model. We look into a more advance state of the art methods in classification task.

1.4 Deep Learning

Deep learning is the current state of the art machine learning approach. Deep learning has become famous in the recent years and it has become more popular in Image domain. However, it has also showed promise in Speech and text domain. It was introduced to close the difference between Artificial intelligence and Machine learning. Deep learning is not a new approach, It is nothing but a neural networks with many layers. The main reason that Deep learning is able to beat the base line models is because of the computation power. We have dedicated hardwares which can make the computation very fast. Deep learning as the name says it is a bunch of layers with non-linear activation function at the end of each layer. The layer can include convolution , fully connected layer, pooling layer. The more detailed explanation on deep learning can be explored in convolutional neural networks website [2]:

1. **LeNet** [7]: LeNet was one the first deep learning model with less than 7 layers. It was introduced by Yan LeCun and it was used to recognize digits in the zip

[2] http://cs231n.github.io/convolutional-networks/

code. This was the first known convolutional neural networks. Many more state of the art models use more than 10 layers to get a standard accuracy.

2. **AlexNet** [8]: AlexNet was another deep learning architecture which won the Imagenet competition in 2012. The AlexNet Architecture is another feed forward deep learning architecture with the collection of convolutional layers,pooling layers,fully connected layers with non-linear activation at the end of each layer. AlexNet showed the world that deep learning can be replaced with the existing machine learning models to get human accuracy performance. AlexNet has 8 layers and the model was trained on GPU.

3. **GoogLeNet** [9]: GoogleNet, was developed by the researches at Google to solve the ImageNet classification problem and they won the competition. The main reason for having a better accuracy than others was using a new architecture was Inception model which we will look in much detail in the upcoming sections. The inception has small convolution filters so that it can capture much small details clearly. we will be using GoogleNet to solve surgical tool classification.

4. **VGGNet** [11]: VGGNet, was developed to solve Imagenet classification problem and it won the competition in 2014 (ILSVRC 2014). The two main reason for this state of the art method was using consistent filter size and many convolutional and pooling layers. We will be using VGGNet for our surgical tool classification.

5. **ResNet** [12]: ResNet (Residual Network) won ImageNet Large-Scale Visual Recognition Challenge 2015 (ILSVRC 2015). They used skip connections and batch normalization which was promising and the architecture does not have fully connected layers at the end of the network. The ResNet was developed by Microsoft researchers and it has more than 150 layers.

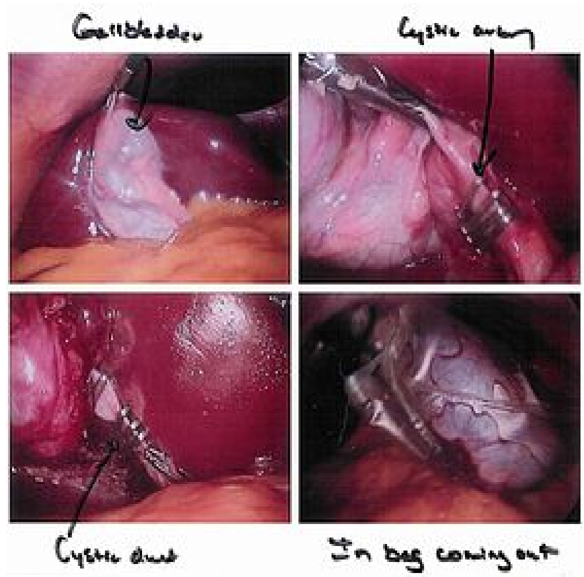


Figure 1.2: Laparascopic image of Cholecystectomy

Figure 1.3: Laparascopic stomach surgery

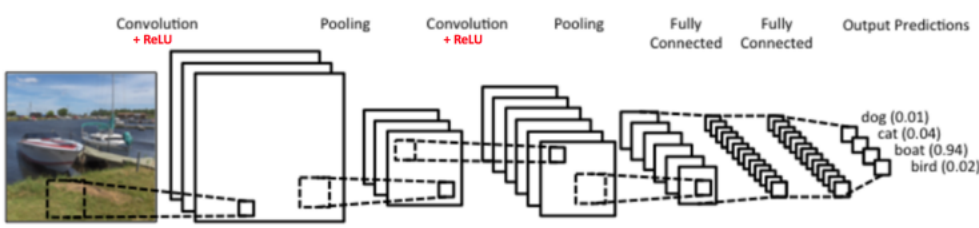


Figure 1.4: ConvNet Architecture

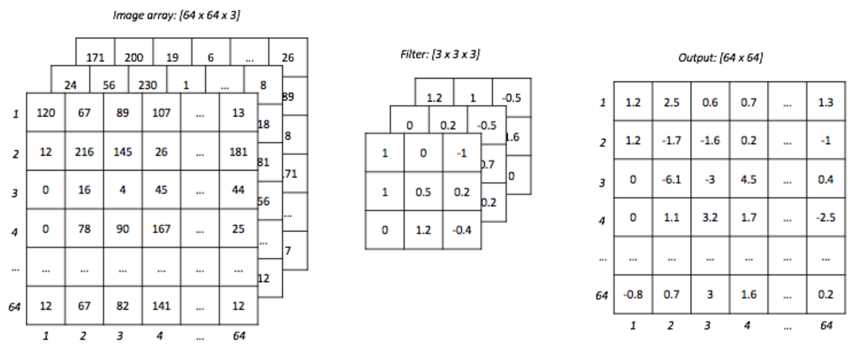


Figure 1.5: Convolution operation

Figure 1.4 [3] illustrates an architecture for ConvNet. Convolutional layer,pooling layer and fully connected layer are the main features of ConvNet.

1. CONVOLUTION: The convolution is nothing but capturing the features of a given image. It can capture shape,edges of the image. This is illustrated in Figure 1.5.
2. POOLING: Pooling layer reduces the size of the input feature keeping the features preserved, as shown in Figure 1.6.
3. FULLY CONNECTED: Fully connected layer combines the weights of previous layer to the next layer. This layer helps to estimate class scores.

[3]https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

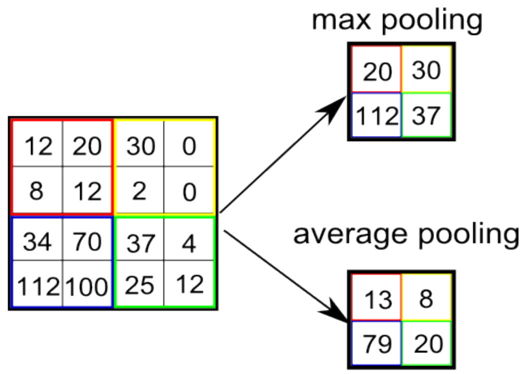


Figure 1.6: Pooling operation

1.5 Goal of Thesis

The goal of this thesis is to build a end to end deep learning pipeline and use ensemble methods to beat the other existing models. We use VggNet and GoogleNet to ensemble and produce a new state of the art results. The classification pipeline consists of four different stages:

1. Video to Image conversion
2. Real time data augmentation
3. Train VGGNet and GoogleNet Deep ConvNet model
4. Ensemble methods

Traditional pipeline does not use the two main state of the art methods and it comprises of computer vision based methods which we will see in classical method section. The data is sponsored by MICCAI'16 [4] grand challenge. Figure 1.7 depicts an overview of the proposed multi-label classification system for identifying surgical tools.

[4] http://camma.u-strasbg.fr/m2cai2016/

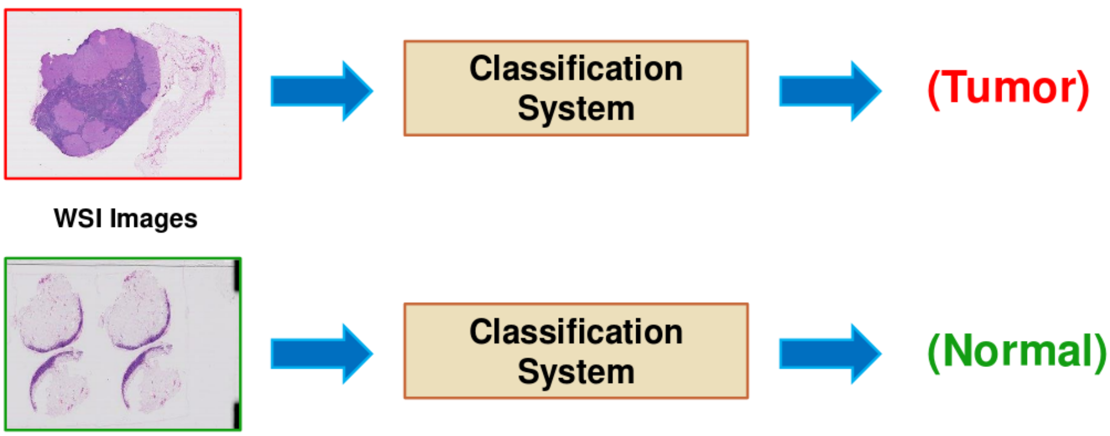


Figure 1.7: Surgical tool Classification System

CHAPTER 2

CLASSICAL APPROACHES

2.1 Image analysis based methods

Since the start of Computer Assisted Intervention, there has been papers and implementation methods to close the gap between manual intervention and automated intervention. These baseline methods [21] involve the use of color normalization, line detection, angle detection and feature extraction [22]. Detailed architecture of these methods is illustrated in Figure 2.1. In this section, we will explain each stage of the classical method in detail.

2.1.1 Gray scale conversion

The images that we get are RGB image and the classical methods use Ostu method [22] to convert to binary image and used to detect the lines in the image which in other words are the tools. In order to do the Ostu method we must first convert the RGB image into Gray scale image or in other words we must convert the three channel image into a single channel image.

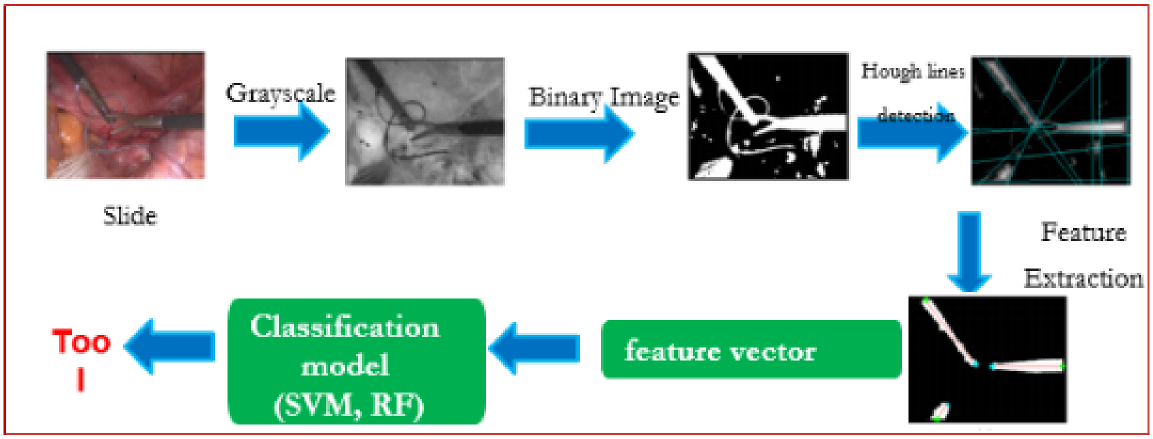


Figure 2.1: Classical method architecture

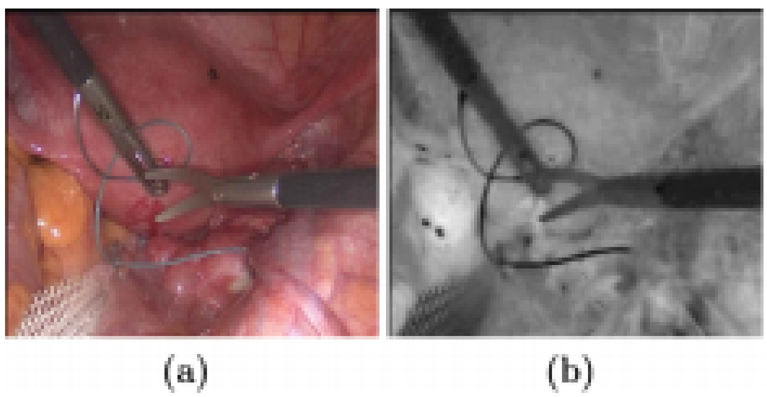


Figure 2.2: An example of converting RGB to Grayscale (a) Source Image (b) Target Image

2.1.2 Binary Image conversion and Otsu method

The second step include Binary Image conversion where the gray scale image is converted to a binary image with has either 0 or 1. The need for binary image conversion is to find the tools that are of different color with respect to the background. After converting to Binary image we can apply Otsu method [22] to detect surgical tool in the image.This principle holds for our domain where the surgical tools are of different color when compared to the background data. There is a possibility of noise

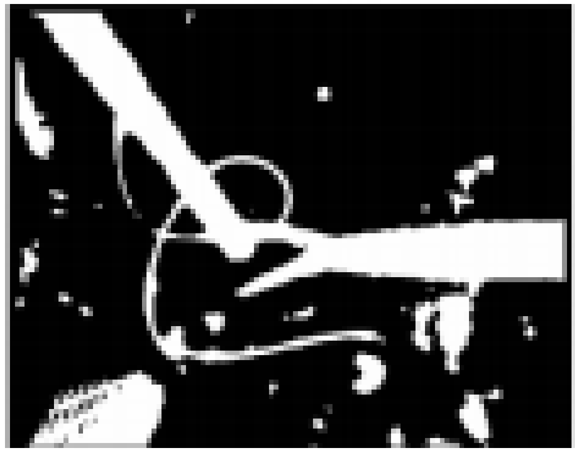


Figure 2.3: An example of Binary Image conversion

while converting to Binary image and this noise is removed by using morphological transformation and in particular using erosion step and using a cross shaped kernel. The edges are identified using canny edge detector. We can see that the tool detection does not take place completely and we have lot of noise in the image. We might need further steps to detect the specific tools in the image.

1. **Binary Image conversion**: Here the grayscale image is converted to binary image.
2. **Otsu method**: Otsus method follows on the basic principle that there are two classes in the image, one class with foreground pixels and another with background pixels. The noise is then removed by erosion. Otsu method is used to separate the classes from the background and since it is a manual threshold we can not guarantee that the otsu method will completely separate tools from the background image.

Figure 2.3 illustrates an example of Binary Image conversion with otsu method.

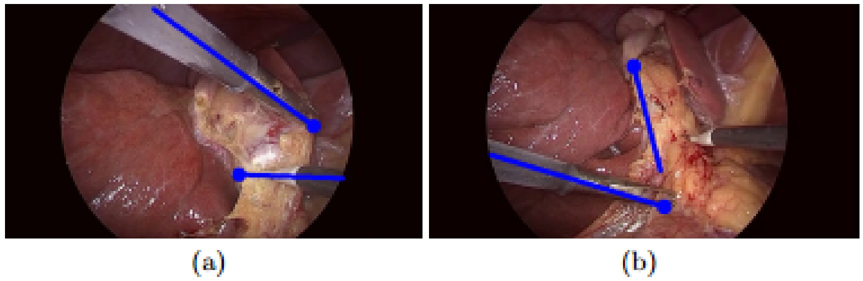


Figure 2.4: An example of Hog lines detection

2.1.3 Feature extraction

The final stage is to extract features and train a classifier on the features to detect the surgical tools. The features are extracted using Hog lines detection [24]. From the image we can see that the features are nothing but the surgical tools that has been extracted from the images. The tools are extracted using the Hog lines detection. The next step is to extract features from the image. The tools information gives information such as 2D tip, angle and lines of the tools in the image. With all these information we can extract features from each and every frame. Our next step we need to be sure that input labeled images are consistent with each other. In order to this we need to have similar feature extraction in all images, features such as size,orientation,color,illumination. The 2D vector between the entry point is given by orientation, size and tip of the detected tool. From this we can do data augmentation such as horizontal,vertical and scaling augmentation. We can see that the size of each image are of different sizes for this the images are re scaled to a fixed image size. The image size was fixed to 256x30 pixels, finally the features are trained using Support Vector Machines [25]. The SVM classifies the features to respective classes.

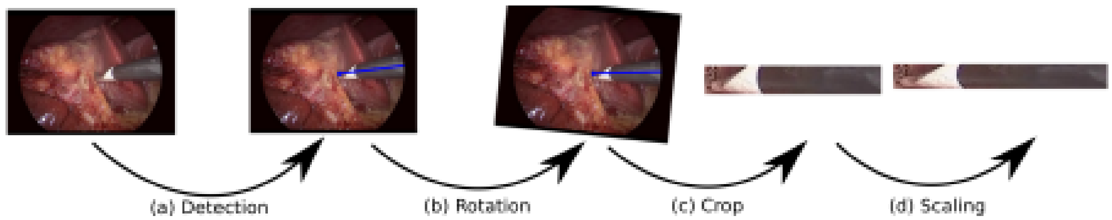


Fig. 3. Tool thumbnail extraction process. (a) Position and orientation detection, (b) Alignment with the tool, (c) Bounding box cropping and (d) Final scaling.

Figure 2.6: An example of data augmentation

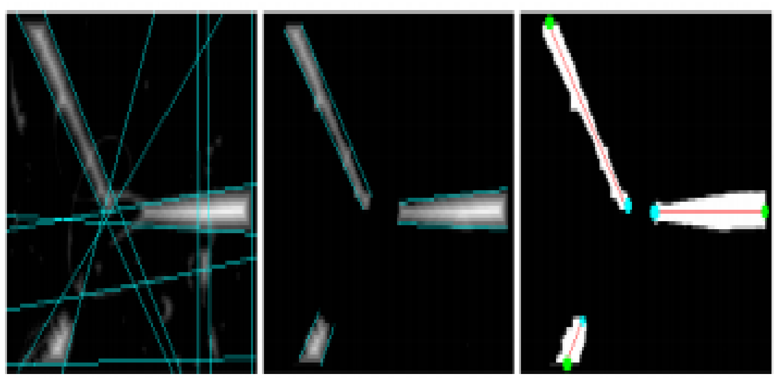


Figure 2.5: (a) Hough lines detection,(b) Pruning and (c) Final image mask with main axis (red), tip (blue) and border (green) for each instrument.

2.2 Challenges

Though classical challenges are simple and much easy to understand the main problem with this classical challenges are the abstraction. The feature extraction method is not generalized and the feature extraction itself is not very efficient. The classical methods use many manual hyper parameters which needs experts advice to tune it. We would see Deep learning approach which is much abstract and much efficient when compared to classical method.

CHAPTER 3

DEEP LEARNING FOR Multi-Label Surgical tool detection

3.1 Surgical Tool classification pipeline

In this thesis, we have focused on the Surgical Tool classification and the important point to note that this is a Multi-label classification, there can be more than one tool present in each video frame and our model has to give probability to each class in the given frame. The model has to identify the probability of each tool present and the probability of each tool not present. Our classification pipeline consists of four stages:

1. Convert Video to Image frames
2. Real time Data Augmentation
3. Train Deep ConvNet models to classify multiple labels in the frame
4. Ensemble methods
5. Post-processing and submit the results

Figure 3.1 depicts Deep learning classification framework.

3.1.1 Convert Video to Image frames

The first step in our pipeline is to prepare the data that would be suitable for classification. The data that is given to us is in the form of videos and the labels are annotated for particular frames. Our first step is to convert videos to images and extract the particular frames which has labels associated with it.

1. **Video to Image conversion**: First step is to convert the raw videos to images at 1 frame per second. The conversion of videos to images was done by Adobe

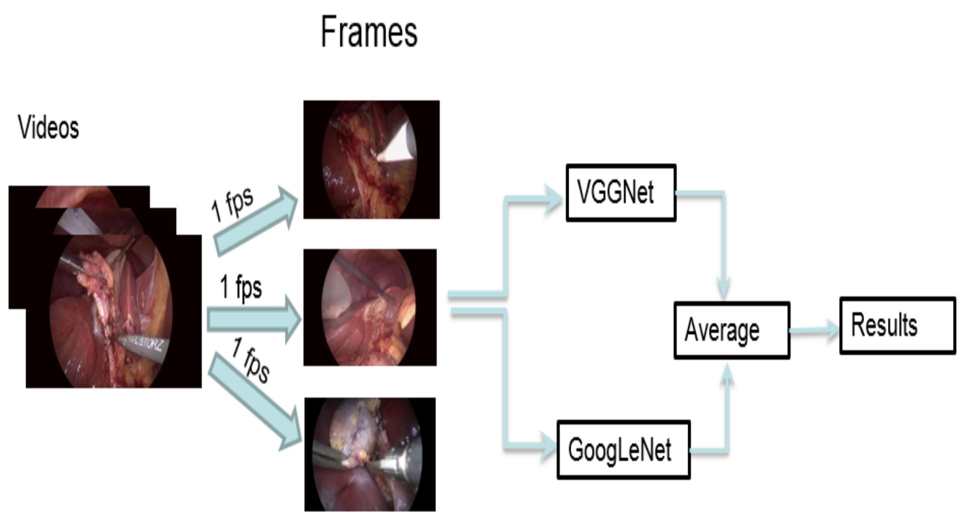


Figure 3.1: Deep learning classification framework [1]

photo shop as they have more sophisticated way to convert videos to images. The videos are converted to 1 frame per second because the annotations are labelled at 1 frame per second.

2. **Annotated Image extraction**: Next step is to extract images that was labelled by the organizers. The image extraction was done for every 25 frames. The frames was extracted using a customized script. The images are then re sized to a default dimension as the model accepts 224x224x3 dimension. The labels are in the form of vector with 1 representing the tool is present and 0 representing the tool is not present as seen in the Figure 3.2

3. **Labels representation**: The labels are represented in the form of one hot encoding where if the tool is present it is marked as 1 and if the tools are not present the tools are marked as 0. The tools are taken in the order as follows: 1) Grasper 2) Hook 3) Clipper 4) Bipolar 5) Irrigator 6) Scissors 7) Specimen bag. For example, If in a given frame Grasper and Bipolar is present, the label

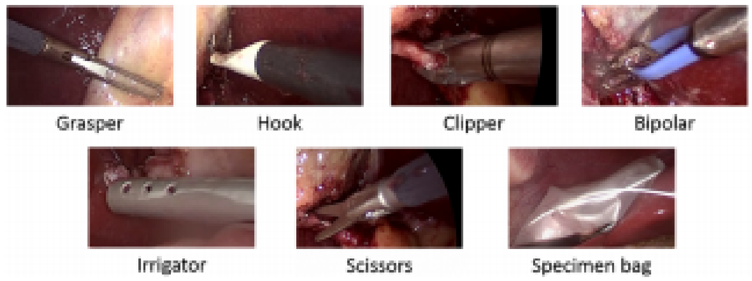


Figure 3.2: labels representation

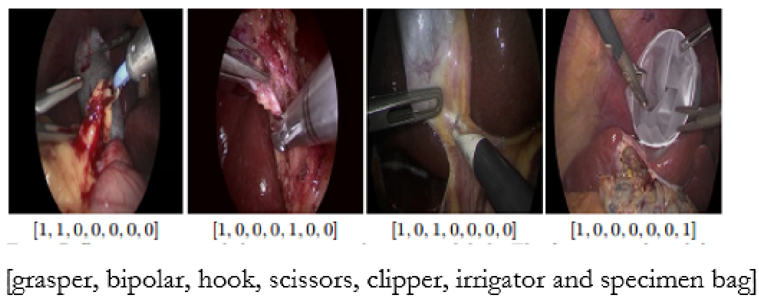


Figure 3.3: labels annotation

would be 1,0,0,1,0,0,0. We can see that our model must be able to take both probabilities if the tools are present as well as if the tools are not present.

3.1.2 Real time data Augmentation

This section explains the data augmentation that was performed. We will discuss the need for data augmentation and the important data augmentation techniques that is mostly used on the images. The total amount of data was around 24000 and we know that for training a deep learning model we need more images. We can see the distribution of surgical tools in the entire data. We can see the distribution of surgical tools in the figure 3.4

Tool numbers

Index	T1	T2	T3	T4	T5	T6	T7
Number	10967	635	14130	411	878	953	1504

Figure 3.4: Tools distribution

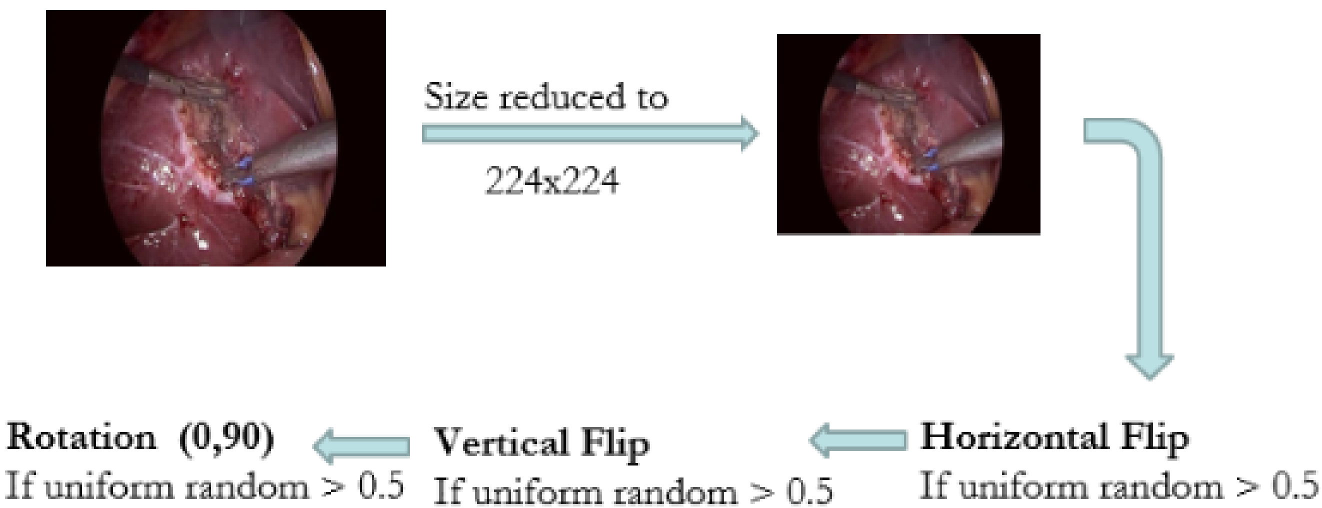


Figure 3.5: Data augmentation techniques

3.1.3 Deep ConvNet Surgical Tool classification

Convolutional neural networks are the most promising field in machine learning. The field became popular after the famous ImageNet competition in which the competition was won by using a deep learning model in 2012 and which improved its accuracy drastically. The ImageNet competition before 2012 was conquered by computer vision techniques which extracts features and then trains a SVM classifier on the extracted features. Deep learning is a more generalized classifier which made it to perform better than the rest. We can see that the Deep learning model is an end to end pipeline and it takes into the account that the features for a particular problem can be learned by itself rather than the user feeding the features to it. The

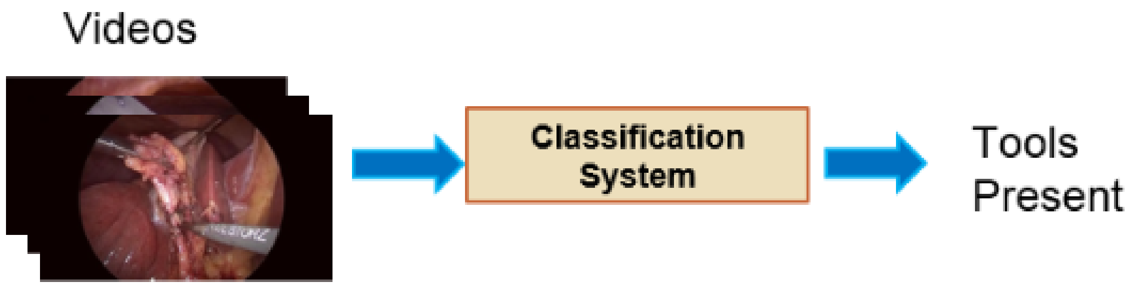


Figure 3.6: Classification pipeline

deep learning model consists of millions of parameters that has to be trained and the only way to train a millions of parameters in a reasonable time is to have a fast computation power. The ConvNet now use GPU's to train their model. The reason of using GPU is because of the multi-processing cores. The Convolutions are performed on the images to extract different features based on the loss we encounter by training the data.

3.1.4 VGGNet

We train two different Deep learning models which are the state of the art. The first one would be VGGNet[3] which won the ImageNet competition in 2014. The VGGNet has 19 layers and layers are comprised of convolutional layer,pooling layer,fully connected layer. Each layer have been discussed already so we will look into how VGGNet is modelled. The VGGNet has over million parameters so the batch size should be kept small. We will look into the hyper parameters that were used in this model from the Figure 3.7. We can see that the learning rate is kept very low because we are not using any pre trained weights from the already built VGGNet from some other data. We use Activation function as Leaky Relu[2]. The reason for using Leaky ReLU is to have a smooth non-linear function rather than the sharp

VGG19 Net

Hyper Parameters	Values
Learning Rate	0.00001
Pre-trained Weights	No
Activation Function	Leaky-ReLU
Loss Function	Sigmoid
Batch-Normalization	Yes
Batch Size	32
Max Epochs	1000
Cross-Validation	5-fold

Figure 3.7: VGGNet hyper parameters

non-linear function. The loss function used is sigmoid rather than a typical softmax because the in any given frame there can be more than one tool so the sigmoid would make as the best choice for us rather than softmax. We use small batch size and the reason for this that the VGGNet has more parameters to learn. The model is trained for 1000 epochs and the model is tested with 5 fold cross validation.

1. GPU: 4 x 12 GB NVidia K40
2. CPU: 3.4GHz Intel core i7 4770
3. HDD: 5 TB
4. RAM: 16 GB DDR4

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 3.8: VGGNet architecture [3]

3.1.5 GoogleNet

We will use another state of the art model called GoogleNet[4]. GoogleNet was developed by Google research fellows and won the ImageNet competition in 2015. The main difference between VGGNet and GoogleNet is the Inception architecture where the inception architecture has many small convolutional filters where the filters capture small features in the images. The GoogleNet hyper parameters can been seen in the Figure 3.9. The hyper parameters have the same learning rate as VGGNet and uses the pre trained weights from the ImageNet model. The GoogleNet used for this competition uses sigmoid rather than softmax. GoogleNet uses Batch-Normalization[5] same as VGGNet model. The batch size is larger than the VGGNet because the number of parameters is less than VGGNet so the computation time is less than VGGNet. GoogleNet has more layers than VGGNet, GoogleNet has 36 layers and each layers have many small convolutional filters. The parameters are trained in a fast way because of the less parameters and more layers. The inception can be altered in any way and can be added with more layers. We have used the standard inception-v3 model with 36 layers. The data has been trained with 5 fold cross validation because 5 fold cross validation is the standard way of finding whether the model is over fitted or not.

Google Net

Hyper Parameters	Values
Learning Rate	0.00001
Pre-trained Weights	Yes
Activation Function	Leaky-ReLU
Loss Function	Sigmoid
Batch-Normalization	Yes
Batch Size	64
Max Epochs	1000
Cross-Validation	5-Fold

Figure 3.9: GoogleNet hyper parameters

3.1.6 Ensemble Methods

After we train the models we need to somehow combine the results of the trained models and produce the best results. The ensemble can be done by weighted averaging and we have tried with many weighted parameters and we found out that the weight parameter is 1 or in other words we do simple averaging of all models. The results of 10 models with random weights in each model gave a relatively low results rather than the equal weights. We have trained 5 models for VggNet and 5

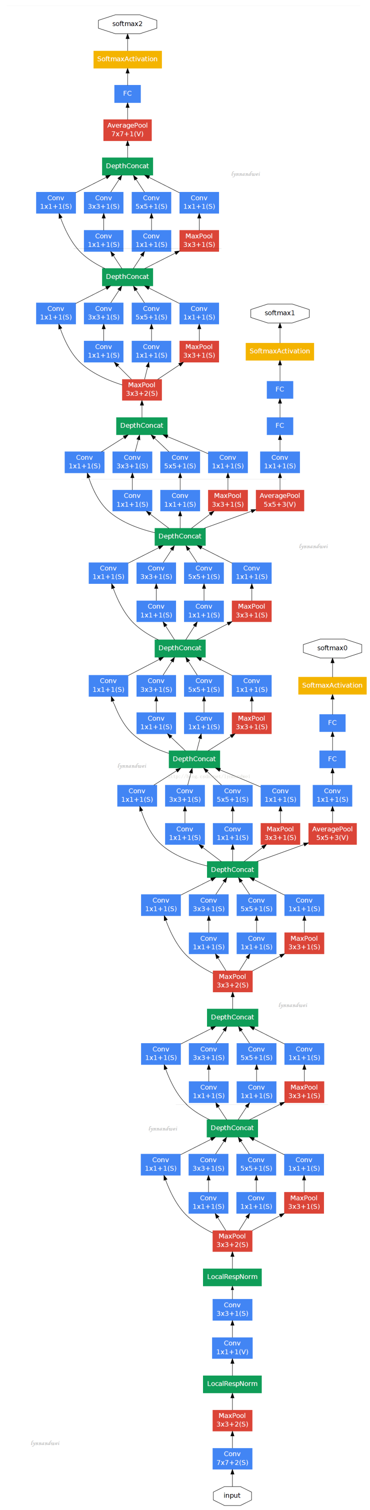


Figure 3.10: Inception-V3 architecture

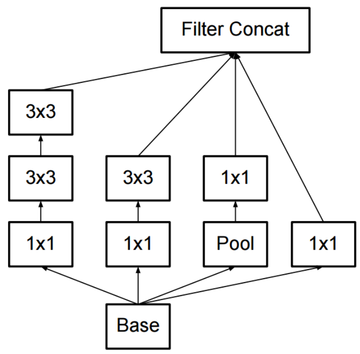


Figure 3.11: Inception modules where each 5 x 5 convolution is replaced by two 3 x 3 convolution [10]

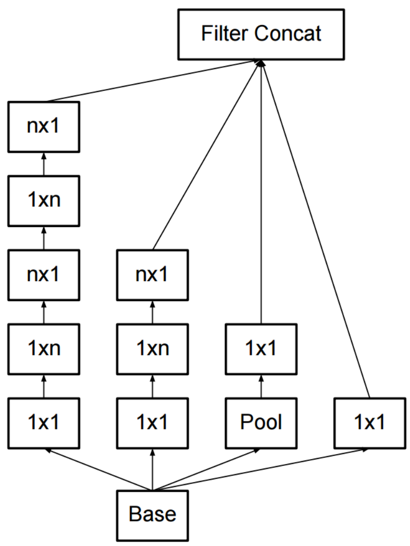


Figure 3.12: Inception modules after the factorization of the n x n convolutions. We chose n = 7 for the 17 17 grid [10]

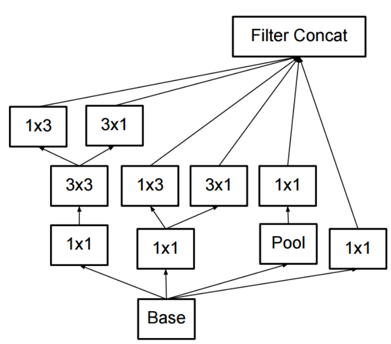


Figure 3.13: Inception modules with expanded the filter bank outputs. This architecture is used on the coarsest (8 x 8) grids to promote high dimensional representations [10]

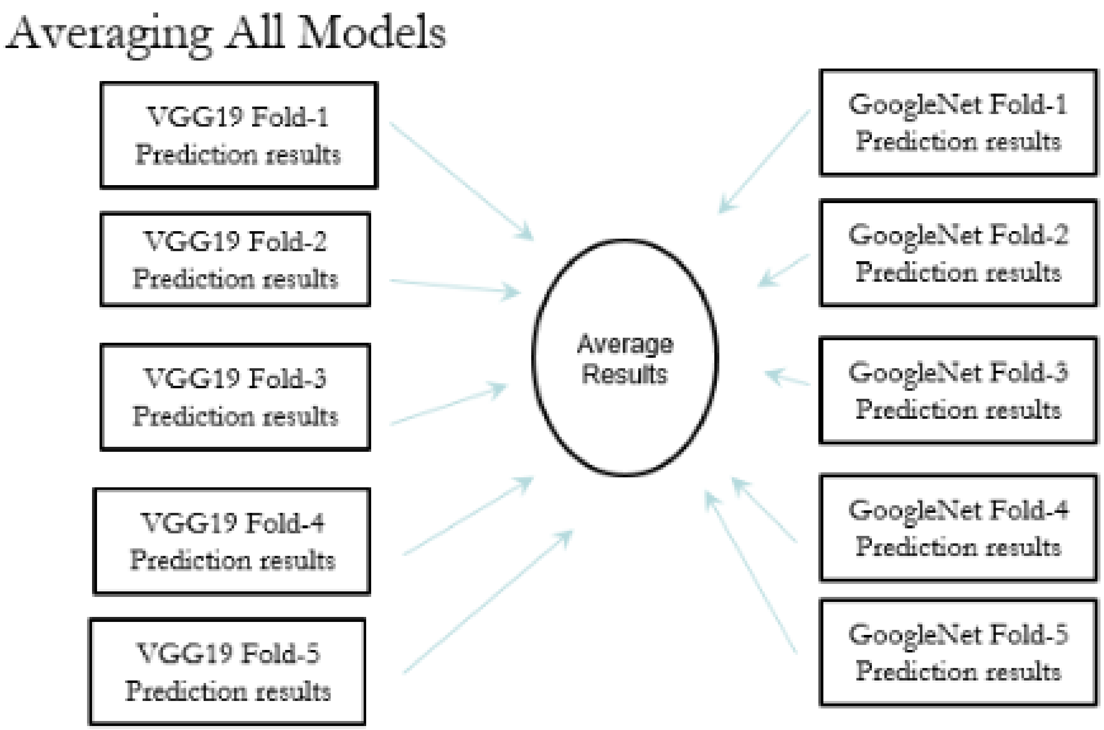


Figure 3.14: Ensemble methods

models for GoogleNet as from the figure 3.1.6 . We average the results of the output that was extracted by the 10 models respectively.

CHAPTER 4

EXPERIMENTAL RESULTS

4.1 Dataset

For this thesis we used the dataset provide by MICCAI 2016 [5]. The dataset consists of 10 training videos and 5 testing videos. Each training video runs at 25 frames per second and the testing videos runs at 1 frame per second. The labels are annotated at every frame per second for certain frames. The frames are labelled in a separate text file and the tools are given with the label 1 or 0. The video frames are extracted for every 25 frames since the labels are annotated at every 25 frames. We have already discussed how we convert the videos to image frames so we will look into the the tools that was used to convert videos to images. The images are of different sizes and it is resized to standard size of 224x224x3. The re sized images are fixed because the model input size. After extracting the images, the total image size is around 24000 and the images are further expanded by doing real time data augmentation. The real time means the data is augmented during run time and the data is augmented based on uniform random probability.

Source	**Training**	**Test**
videos	10	5
images	24000	18000
Total after augmentation	120000	18000

Table 4.1: MICCAI surgical tool dataset

Figure 4.1: Miccai Challenge

4.2 Experimental setup

The experiments were setup in the SMILE Lab and the hardware configurations of the system were:

1. CPU:3.4GHz Intel core i7 4770
2. RAM: 16GB DDR4
3. 18 GB of NVidia K40 GPU
4. 12 GB of NVidia Ge Force Titan X GPU

The software requirements were:

1. OS: Ubuntu 14.04
2. Programming Languages: Python 2.7
3. Deep Learning libraries : Tensorflow (v0.12.1)
4. Support libraries: Sci-Kit, NumPy
5. Video-image conversion: Adobe photoshop
6. pre-trained models: Caffe-zoo

4.3 Evaluation metrics

4.3.1 Hamming distance

Hamming distance is the measure which says how different are the two vectors. Hamming loss takes into account of both tools present and not present. Hamming loss takes the intersection over union of the predicted vector and the ground truth vector. Here the vector is considered as the the labels with 1 or 0. For example, [1,1,0,0,0,0] is a label with 2 tools present and 4 tools not present. So taking the intersection over union of predicted and the ground truth gives the probability of tools present and not present as well. The hamming loss is found to be performing better than any other loss for the multi-label classification problem. The Evaluation metric is very important for the model to measure its accuracy. The accuracy is the measure of whether tools are present or not.

4.4 Results

4.4.1 Training and validation

The training took 4 days with 1000 epochs for each model and we used 5 cross validation. For cross validation, we use 80 percent of the data as training and 20 percent of the data as validation and for next set of validation we take another set of the validation. we can see the validation results from the figure 4.3. We can see that the results of one fold in VGGNet is low and the results of the same fold in Googlenet is high. So the ensemble method would make the score better. The results of each fold is non evenly distributed and ensemble method of these two models would result in better results. Training with multiple gpu would make the compute time faster but we restricted with 2 gpu since we had hardware constraints.

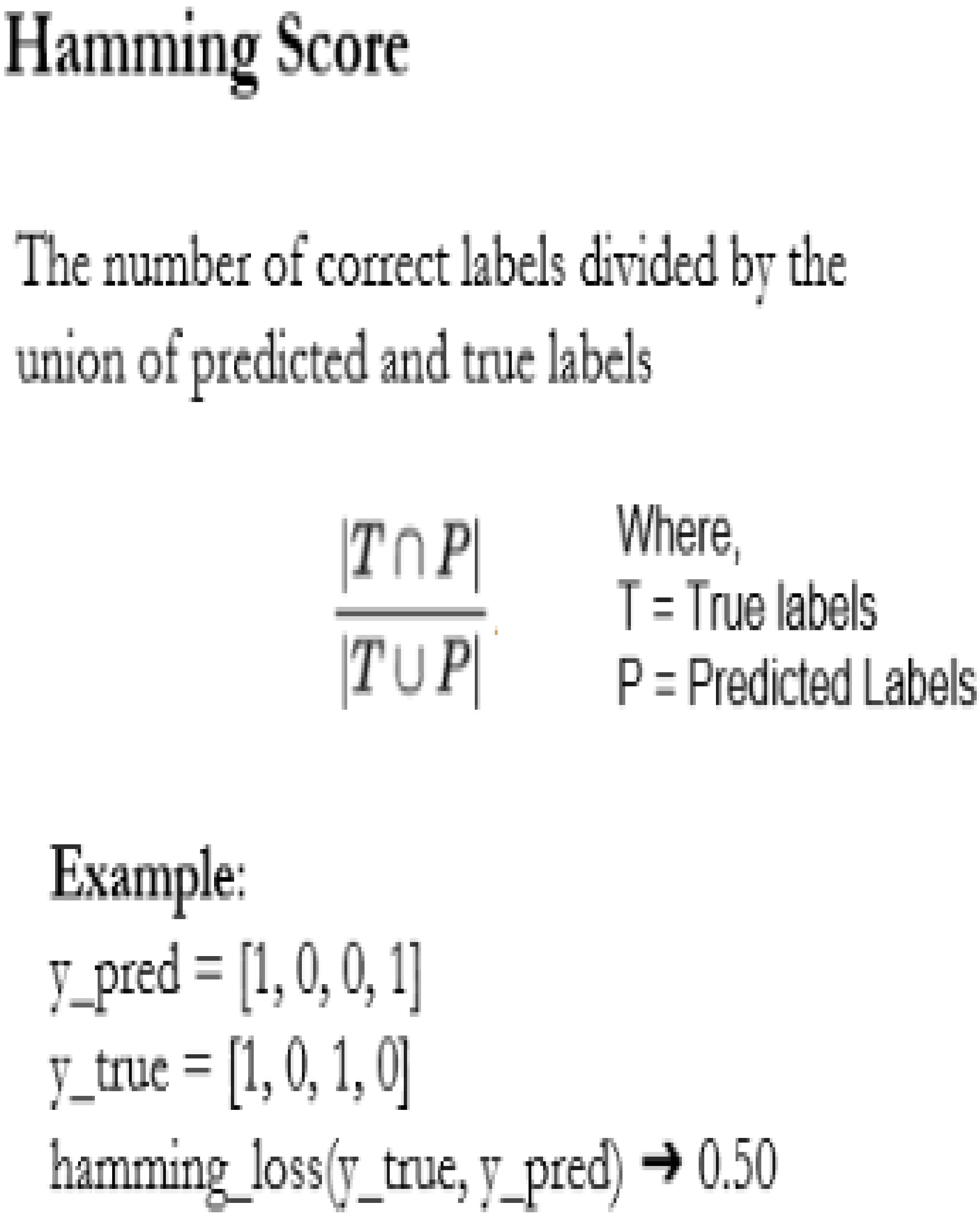


Figure 4.2: Evaluation metric

4.4.2 results

VGGNet Validation Score

Fold-1	Fold-2	Fold-3	Fold-4	Fold-5
78.5	76.2	80	81	72.5

GoogLeNet Validation Score

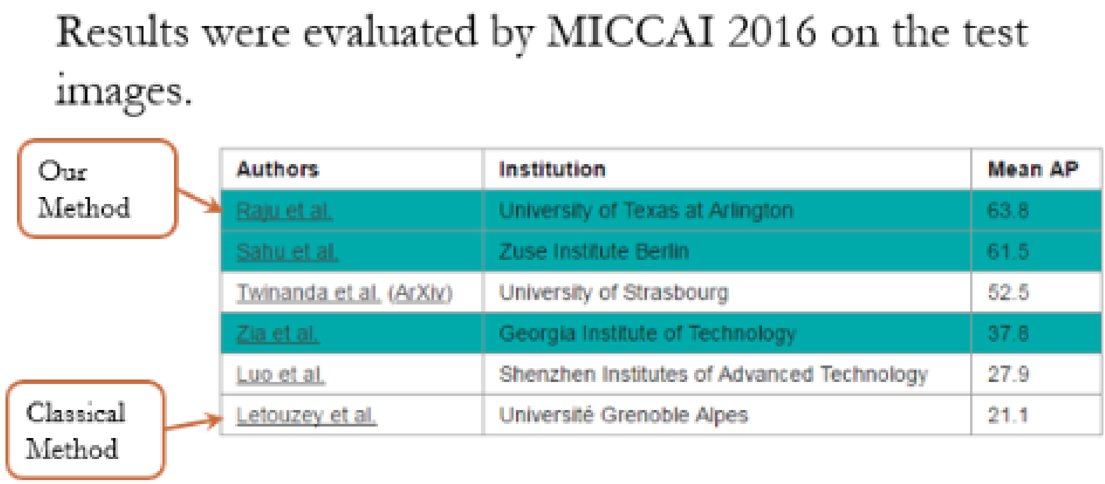
Fold-1	Fold-2	Fold-3	Fold-4	Fold-5
83	81.7	79	70.4	82

Ensemble Validation Score : 78.3%

Figure 4.3: Training results

4.5 Ensemble method

As seen in Figure 4.3, The ensemble method increases the validation results when compared to individual model results. The validation results produces the score with 78.3%. The validation results yields a better score and the score seems to be much better than the base line method. Our model yields 3 times better than the base line method. We first tried with weighted averaging where the weights are kept random so results are randomly sampled and compared with other results. The simple averaging without any weights produced a better results than the weighted results.



Results were evaluated by MICCAI 2016 on the test images.

Authors	Institution	Mean AP
Raju et al.	University of Texas at Arlington	63.8
Sahu et al.	Zuse Institute Berlin	61.5
Twinanda et al. (ArXiv)	University of Strasbourg	52.5
Zia et al.	Georgia Institute of Technology	37.8
Luo et al.	Shenzhen Institutes of Advanced Technology	27.9
Letouzey et al.	Université Grenoble Alpes	21.1

Figure 4.4: Test results

The Test results was evaluated the MICCAI challenge sponsors and they used Mean Average Precision and the results are shown in 4.4. The reason for using Mean Average Precision is nothing but the Mean of all Average precisions. The MaP is another evaluation metric as Hamming distance. The ensemble mehtod can be further improved with more models added to it. The another state of the art deep learning model such as ResNet[6],AlexNet can also be added. The model with more layers generally produce a better results so adding more models would produce more accuracy. The ensemble methods work better in most of the cases as they try to overcome the drawback in one model with the other.

4.6 Further improvements

Further improvements can be done using the advantage of sequence learning. The improvement provided by Long Term Short Memory [7] recently shows that the

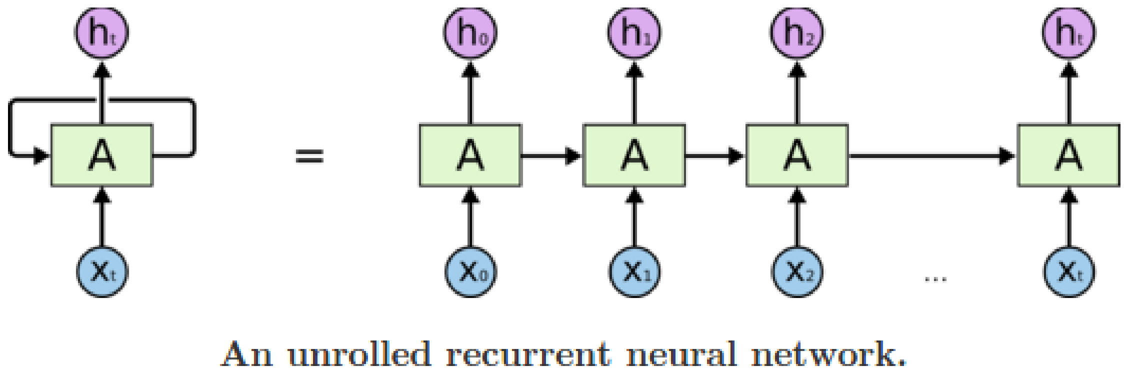


Figure 4.5: LSTM example

prediction of particular time can be done using the model trained on the time until that period. LSTMs with many layers found very promising and it can be applied to surgical Tool detection because of the fact that the tools in the current frame is based on the tools in the previous frame. LSTM can be seen in this figure **??**

4.7 Results comparison

Figure 4.4 compares result obtained by the proposed method and it is compared with other 5 models. We can see that the base line model is 3 times worse than our model's prediction. The results are submitted to MICCAI and our model won the first prize in the competition.

CHAPTER 5

CONCLUSION AND FUTURE WORK

The goal for this thesis is to build the deep learning model for automatically classifying the surgical Tools in a Laprascopic surgery. The future of Computer Assisted Intervention would be to fully automate surgery process and this challenge would help solving the main problem much soon. The main challenges we came through while solving the problem was the variance in images and the less data set size. We first saw the classical methods where there were series of stages starting with feature extraction and then training a classifier to find the tools matching the features in the test image. The feature extraction stage includes the standard computer vision methods such as Otsu method for gray scale to binary image conversion, Hog lines detection and many more methods. The feature extraction for each problem is different and it needs an expert person to set the manual hyper parameters for each and every problem. The classical methods also have problem with using a weak classifier when compared to Deep learning classifier so the classical methods do not provide a good way to classify the objects as we expected. The Deep learning model performs better when we increase the layers and we experimented with the two state of the art methods one with VGGNet and another with GoogleNet. The reasons for using VGGNet and GoogleNet is that the model won the ImageNet competition in 2014 and 2015 respectively. We further tried with hyper parameter customization and found the best hyper parameter for our models. We further found that if we combine both models and the ensemble results would produce better results than the single model. The model is trained with cross validation and with 5 fold. The 5 fold cross validation is the standard way of measuring over fitting model. The ensemble model with simple averaging was found to work better than the other models. The test results was evaluated by MICCAI organizers and the results was published during the workshop. Our model received the first prize and it was the deep learning model that was the reason behind it. We would like to explore LSTMs on surgical tools as we know that the surgical tool on the current frame is based on the previous frame. LSTMs with modified kernels such as GRU would make the performance much better.

REFERENCES

[1] Sheng Wang, Ashwin Raju, Junzhou Huang “A Multi-Label classification Deep learning Method for Surgical Tool presence detection in laparoscopic videos“ ISBI 2017

[2] Bing Xu, Naiyan Wang, Tianqi Chen, Mu Li “Empirical Evaluation of Rectified Activations in Convolutional Network “

[3] Karen Simonyan Andrew Zisserman “ VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION “

[4] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich “Going Deeper with Convolutions“

[5] Sergey Ioffe, Christian Szegedy “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift“

[6] “http://camma.u-strasbg.fr/m2cai2016/index.php/program-challenge/“

[7] Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun “ Deep Residual Learning for Image Recognition“

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”, In NIPS, 2012.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. “Going deeper with convolutions”, in CVPR, pp. 1-9, 2015.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. “Rethinking the inception architecture for computer vision”, arXiv preprint arXiv:1512.00567, 2015.

[11] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”, CoRR, vol. abs/1409.1556, 2014.

[12] He K, Zhang X, Ren S, Sun J. “Deep Residual Learning for Image Recognition”, CVPR 2015.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. “Imagenet large scale visual recognition challenge”, International Journal of Computer Vision, 115(3):211252, 2015.

[14] N. Otsu. “A Threshold Selection Method from Gray-level Histograms”, IEEE Transactions on Systems, Man and Cybernetics, 9(1):6266, 1979.

[15] Raphael Sznitman, Carlos Becker, and Pascal Fua, “ Fast part-based classification for instrument detection in minimally invasive surgery,“ in International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, 2014, pp. 692699.

[16] Huey Nee Lim, Mohd Yusoff Mashor, Nadiatun Zawiyah Supardi, Rosline Hassan. “Color and Morphological Based Techniques on White Blood Cells Segmentation”, in Proc. IEEE 2nd Intl. Conf. on Biomed. Engineering, 2015.

[17] S. Ioffe and C. Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, CoRR, vol. abs/1502.03167, 2015.

[18] Nitish Srivastava , Geoffrey Hinton , Alex Krizhevsky , Ilya Sutskever , Ruslan Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting”, The Journal of Machine Learning Research, v.15 n.1, p.1929-1958, January 2014

[19] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. "Mitosis detection in breast cancer histology images with deep neural networks", In Medical Image Computing and Computer-Assisted InterventionMICCAI 2013, pages 411418. Springer, 2013.

[20] Antoine Letouzey1, Marion Decrouez2, Anthony Agustinos1,Sandrine Voros "Instruments Localisation and Identification for Laparoscopic Surgeries"

[21] Otsu, N.: "A threshold selection method from gray-level histograms. Automatica (1975) "

[22] Agustinos, A., Voros, S.: " 2D/3D Real-Time Tracking of Surgical Instruments Based on Endoscopic Image Processing. Computer-Assisted and Robotic Endoscopy" Second International Workshop, (CARE 2015)

[23] Dalal, N., Triggs, B.: "Histograms of Oriented Gradients for Human Detection".International Conference on Computer Vision Pattern Recognition (CVPR 2005).

[24] Chih-Chung, C., Chih-Jen, L.: "LIBSVM: A Library for Support Vector Machines." National Taiwan University, Taipei, Taiwan (2001)

[25] Kevin Cleary, Ho Young Chung, and Seong K Mun, "Or2020 workshop overview: operating room of the future," in International Congress Series. Elsevier, 2004, vol. 1268, pp. 847852.

[26] Abdel-Hamid, O., Mohamed, A., Jiang, H., and G. Penn. "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition", Proc.ICASSP, 2012.

[27] Darko Katic, Anna-Laura Wekerle, Fabian G artner, Hannes Kenngott, Beat Peter Muller-Stich, R udiger Dillmann, and Stefanie Speidel, "Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-

aware assistance," in International Conference on Information Processing in Computer-Assisted Interventions. Springer, 2014, pp.158167

[28] Zheng Xu, Junzhou Huang. "Detecting 10,000 Cells in One Second", In Proc. of the 19th Annual International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'16, Athens, Greece, October 2016.

[29] Sheng Wang, Jiawen Yao, Zheng Xu, Junzhou Huang. "Subtype Cell Detection with an Accelerated Deep Convolution Neural Network", In Proc. of the 19th Annual International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'16, Athens, Greece, October 2016.

[30] Jiawen Yao, Sheng Wang, Xinliang Zhu, Junzhou Huang. "Clinical Imaging Biomarker Discovery for Survival Prediction on Lung Cancer Imaging Genetic Data", In Proc. of the 19th Annual International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'16, Athens, Greece, October 2016.

[31] Zheng Xu, Junzhou Huang. "Efficient Lung Cancer Cell Detection with Deep Convolution Neural Network", 1st International Workshop on Patch-based Techniques in Medical Imaging, PMI'15, Munich, Germany, October 2015.

[32] Xinliang Zhu, Jiawen Yao, Feiyun Zhu and Junzhou Huang. "WSISA: Making Survival Prediction from Whole Slide Pathology Images", CVPR 2017.

[33] R. C. Gonzalez and R. E. Woods. "Digital Image Processing", Upper Saddle River, NJ, USA: Pearson, 2008.

[34] Zhu, Feiyun and Wang, Ying and Xiang, Shiming and Fan, Bin and Pan, Chunhong. "Structured sparse method for hyperspectral unmixing", ISPRS Journal of Photogrammetry and Remote Sensing, 88:101-118, 2014.

[35] Zhu, Feiyun and Liao, Peng and Zhu, Xinliang and Yao, Yaowen and Huang, Junzhou. "Cohesion-based Online Actor-Critic Reinforcement Learning for mHealth Intervention", arXiv:1703.10039, 2017.

[36] Li, Haichang and Wang, Ying and Xiang, Shiming and Duan, Jiangyong and Zhu, Feiyun and Pan, Chunhong. "A label propagation method using spatial-spectral consistency for hyperspectral image classification", International Journal of Remote Sensing, 37(1):191-211, 2016.

[37] Cheng, Guangliang and Wang, Ying and Zhu, Feiyun and Pan, Chunhong. "Road extraction via adaptive graph cuts with multiple features", Image Processing (ICIP), IEEE International Conference on, 3962-3966, 2015.

[38] G. H. Lyman, A. E. Giuliano, M. R. Somerfield, A. B. Benson, D. C. Bodurka, H. J. Burstein, A. J. Cochran, H. S. Cody, S. B. Edge, S. Galper, et al. "American society of clinical oncology guideline recommendations for sentinel lymph node biopsy in early-stage breast cancer", Journal of Clinical Oncology, 23(30):77037720, 2005.

[39] G. H. Lyman, S. Temin, S. B. Edge, L. A. Newman, R. R. Turner, D. L. Weaver, A. B. Benson, L. D. Bosserman, H. J. Burstein, H. Cody, et al. "Sentinel lymph node biopsy for patients with early-stage breast cancer: American society of clinical oncology clinical practice guideline update", Journal of Clinical Oncology, 32(13):13651383, 2014.

[40] Feiyun Zhu and Bin Fan and Xinliang Zhu and Ying Wang and Shiming Xiang and Chunhong Pan. "10,000+ Times Accelerated Robust Subset Selection ARSS", Proc. Assoc. Adv. Artif. Intell. (AAAI), 3217-3224, 2015.

[41] Digital pathology, In Wikipedia. Retrieved April 25, 2017, from https://en.wikipedia.org/wiki/Digital_pathology

[42] Cheng, Guangliang and Zhu, Feiyun and Xiang, Shiming and Pan, Chunhong. "Road Centerline Extraction via Semisupervised Segmentation and Multidirection Nonmaximum Suppression", IEEE Geoscience and Remote Sensing Letters, 13(4):545-549, 2016.

[43] Cheng, Guangliang and Wang, Ying and Gong, Yongchao and Zhu, Feiyun and Pan, Chunhong. "Urban road extraction via graph cuts based probability propagation", 2014 IEEE International Conference on Image Processing (ICIP), 5072-5076, 2014.

BIOGRAPHICAL STATEMENT

Ashwin Raju received his Bachelors in Information Technology in 2013 from SSN college of Engineering, Tamil Nadu, India. After bachelors, he worked in Infosys Limited as a software developer and Business Analyst for 1 year and 8 months. He started his Master in Science degree focusing in Computer Science from 2015 to 2017. His areas of interest include Deep learning, Data Mining, Optimization.