LUNG CANCER DETECTION USING IMAGE PROCESSING ON CT SCAN IMAGES

Leong Shu Ling (26546981)

&

Moana Yamanouchi (27742407)

Supervised by: Dr. Anuja Dharmaratne

Assignment 1 Part 2

FIT 3081 Image Processing

Monash University Malaysia

School of Information Technology

27 April, 2018

Abstract

Lung cancer has the highest mortality rate amongst human beings; hence, early detection is a crucial step to decreasing its percentage and to support early medical treatment. In this paper, various image processing technique will be implemented using Computed Tomography (CT) images to improve the quality of the image to accurately detect and classify between abnormal and normal lung. The proposed system contains three main stages; pre-processing, segmentation and extraction. The proposed method will utilize Median and Gaussian filter for smoothing and Gabor filter for enhancement during the pre-processing stage, Thresholding approach for segmentation and feature extraction based on area and roundness metric for classification purpose.

Contents

1. Introduction

1.1 Background

Cancer is defined as disease involving abnormal cell growth that can begin at any part of the human body. Lung cancer happens when cells at the lung region behave abnormally and form either benign or malicious tumours. The difference between the two is that the former tends to remain in the same place while the latter spreads to other body parts via either bloodstream or lymphatic system, which is referred to as metastasis. Basically, lung cancer can be broadly broken down into two categories based on the appearance of lung cancer cells, which are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). Almost 85% of the lung cancer belongs to the non-small cell lung cancer group, with the remaining 10 to 15% belonging to small cell lung cancer[1] which is mainly caused by smoking. Lung cancer is reported to be the cancer type that has the highest mortality rate among males and females. According to the report given by American Cancer Society[2], it is estimated that about 234,030 new cases and 154,050 deaths will be caused by lung cancer in the year of 2018. Therefore, attention is given to this area in an attempt to increase recovery rate with the help of early and accurate detection. Many researches has conducted on the Computed Tomography (CT) scans to retrieve important information using image processing techniques. For this project, we read up some of the proposed methods and algorithms and choose the most suitable ones in each processing stage to implement a simple lung CT scan classifier.

1.2 Aims and Objectives

CT is known to be the most sensitive imaging modality for pulmonary nodules detection[3] In order to obtain a high quality CT image, it requires a high dosage of radiation which will put the patient's health at risk. Therefore, radiologist must set at low radiation dose, inevitably causing low quality image creating various types of noises, such as salt and pepper, speckle and Gaussian noise, in the CT image thus making the image difficult to detect lung nodule at an early stage. However, it is crucial to detect malignant nodule to provide earlier treatment to reduce the mortality rate.

The aim of this paper is to obtain high quality image at an early stage and to accurately classify between normal and abnormal lung in a CT scan images using MATLAB.

Below are the objectives that we wish to attain at the end of the project:

- Critically analyze various image processing techniques for each stage of pre-processing, segmentation and extraction
- Evaluate which technique is the most suitable and effective for each stage of pre-processing, segmentation and extraction
- Extract CT images from Lung Image Database Consortium Image Collection (LIDC-IDRI) dataset[4]
- Design and develop a system that produces accurate cancer detection results by correctly applying the algorithms in MATLAB

- To make recommendation for further improvement for future scope to give more accurate result for lung cancer detection, if necessary.

2. Literature Review

2.1 Pre-processing

Before performing any kind of operations to extract valuable data and information, images usually undergo a phase known as pre-process. The aim is to improve the quality of image by suppressing or eliminating any unwanted noises and enhancing certain image features for better processing at later stage.

For smoothing, median filter has been used as an effective technique that produces high spatial frequency while maintaining its sharpness and location of the edges[5]. It is a non-linear operation that provides sufficient noise reduction capabilities. Several papers have also combined median filter with other filtering techniques to improve the quality of the CT image. Makaju et.al[6], for example, have applied median filtering and Gaussian filter to remove both salt and pepper as well as speckle noise from grayscale CT scan images.

As for image enhancement, Gabor filter is applied as it is known for its optimal localization properties in spatial and frequency domain[7]. According to the paper written by Chaudhary and Singh[8], they made a comparison between three enhancement algorithms and found out that Gabor filter gave better result compared to Auto Enhancement and Fast Fourier Transform. Several authors have also suggested to Gabor filter based on this comparison [9, 10]. Altarawneh[11] performed similar comparison and obtained same result with FFT giving the worst enhancement percentage.

Miah and Yousuf[12], on the other hand, first converted the image into grayscale and normalized it to 150 x 140 pixels and 200 x 250 pixels. Then, they applied median filter before converting it to binary image. As the final step of pre-processing, they removed unwanted pixels (0s) from the image. They also checked the number of black pixels against their threshold value to differentiate between normal and lung cancer patient to train their neural network.

2.2 Segmentation

Segmentation simplifies and breaks down into constituent regions or objects to allow easier analysis. This is the process that separates foreground objects from the background. Thresholding based method is one of the basic technique that is used to convert grey-level to binary image where each pixel is determined based on the threshold value.

It is reported that compared to Thresholding approach, Watershed Segmentation is more superior[8]. This result is further supported by another paper which stated that Watershed Segmentation has an accuracy of 85.165% compared to Thresholding, which is 81.835%[11]. Also, Makaju et.al[6] proclaimed that Watershed Segmentation can separate and identify touching objects in image. Further enhanced process which 'marks' the object in the image known as Marker-Controlled Watershed Segmentation is also used[9, 10]. For further improvement, Marker-Controlled Watershed with Masking approach has been implemented resulting in earlier detection of lung cancer[7]. It has also given the best performance in segmentation result and running time which was compared with Region Growing and Marker-Controlled Watershed after implementing image enhancement using Gabor filter[13].

Another approach is by converting grayscale image into edge only image. It is then converted into dilated image, followed by filled image. Two lungs are binarized and compared against a threshold value[12].

2.3 Extraction

Feature extraction is the phase where certain important features are extracted from a large set of data to determine the nature of the image. Altarawneh[11] proposed two methods, which are binarization and masking approach to predict probability of lung cancer presence. Binarization basically depends on whether number of black pixels exceeds a certain threshold while masking depends on white connected areas inside region of interest (lungs). It was given that binarization, with a threshold of 17178.48, is superior to masking in both True Acceptance Rate and False Acceptance Rate. In Chaudhary and Singh's paper[8], they aimed to identify the stage of lung cancer, which was they extracted the average intensity of image, area, perimeter and eccentricity as the basis for their classification stage. These features are similar to what Makaju et.al[6] proposed with the addition of centroid and diameter. Miah and Yousuf[12] went a step ahead and extracted 33 features from lung cancer CT scans. They then used these knowledge to train a neural network for classification purpose. However, as for this project, our purpose for now is to only identify the owner of CT scan as normal or lung cancer patient and thus, this approach does not apply to us.

3. Methodology

For the purpose of this project, we are interested in developing a system that takes CT scan images as input and classifies it as lungs belonging to either normal patient or lung cancer patient. The whole system can be divided into several stages, which are image acquisition, pre-processing, segmentation and extraction. Throughout the whole project, MATLAB is used because of its huge database of built-in algorithms for image processing as well as its ability for us to test our algorithms without recompilation. Figure 1 shows the overall system architecture, where the detailed steps involved in each phase is explained in the following subsections.

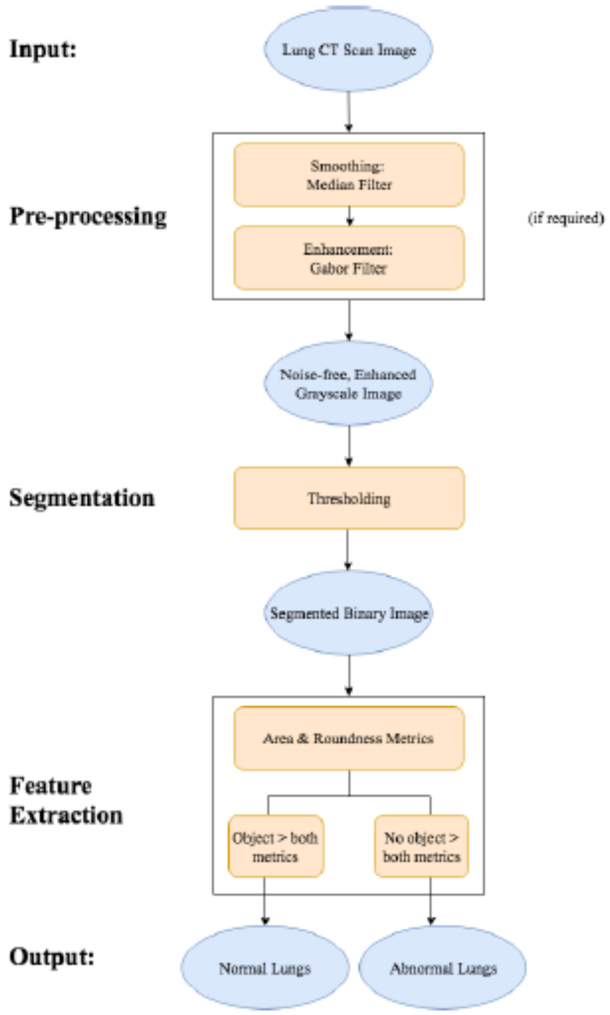

Figure 1

3.1 Image Acquisition

Lung CT scan images which will be used throughout the whole project are obtained from the Lung Image Database Consortium image collection (LIDC-IDRI). All of the images are in DICOM format, with the size of 512 x 512 pixels. DICOM format images are harder to proces[6], thus we have decided to use a software called Horos, which not only allows us to view the DICOM images but also convert these dataset images into JPEG format. These images will then be converted into grayscale images in MATLAB before sending them into the system for further processing. In this project, our dataset consists of a total of 69 CT images, with 30 coming from the normal CT image group and 39 from the cancer CT image group.

3.2 Pre-processing

3.2.1 Smoothing

Firstly, median filter is implemented for noise removal from the image where value of pixel is replaced by the median intensity of the neighbours. This can be mathematically defined as:

$$f'(x, y) = median\{g(s, t)\} \tag{1}$$

where *f'* represents restored image of point *(x,y)* after degradation and restoration process of *f(x,y)*. The median filter computes the median value of image *g(x,y)* in the defined area of *Sxy*.

3.2.2 Enhancement

Gabor filter is a linear filter that can be utilized to extract fundamental frequency of the image due to the frequency response used in the Fourier transform. The 2-D Gabor filter is a product of complex sinusoid and 2-D Gaussian shaped function. To find its impulse response to enhance the image, the two functions are formed into a real number as

$$\psi(x, y) = e^{-(\frac{\alpha^2+\gamma^2 y'^2}{2\sigma^2})} \cos\left(2\pi \frac{x'}{\pi} + \psi\right)$$

$$\text{where } x' = xcos\theta + ysin\theta \text{ and } y' = -xsin\theta + cos\theta \tag{2}$$

where, λ is the sinusoidal factor, θ is the orientation of the normal to the parallel lobes, ψ represents is the phase offset, standard deviation σ of Gaussian function calculates the weighted summation of the surrounding pixels, γ is the spatial aspect ratio.

3.3 Segmentation

Segmentation is applied onto the image to isolate the region of interest for better analysis. Thresholding will be implemented for this project, regardless of the better performing Watershed Segmentation, to prepare for the next stage in our system, which will require objects to be easily identified. Also, Thresholding has faster processing speed and fewer storage space requirement.

Thresholding, as the name suggested, utilizes a threshold value that effectively separates the objects from the background by comparing the pixel intensity. Thus, to find the threshold value, Otsu's method defined the within-class variance as

$$\sigma_w^2 = \omega_0\sigma_0^2 + \omega_1\sigma_1^2 \tag{3}$$

where w is the respective class probabilities computed from a 256-bin histogram separated by a threshold t and σ^2 are variances of the two classes. The t that gives the minimum within-class variance is set to be the final threshold value. As proven by others, this t will also give the maximum between-class variance. For any pixel that has an intensity above the threshold value, it will be regarded as an object pixel, otherwise, it is a background pixel.

Since there are only two classes, 0 and 1, Thresholding will take grayscale CT scan images as input and outputs segmented binary images. For our case, we will use MATLAB's built-in function using "`imbinarize`", which happens to be an algorithm that uses Otsu's method to find the threshold value. To obtain a clean lung-only image for feature extraction, we will apply "`imclearborder`" to clear the image border. As the last step for segmentation, "`bwareafilt`" is used to extract the 5 largest desired objects that are considered as potential cancer nodules and pass on to the next phase.

3.4 Extraction

Observation on multiple CT scan images has led us to come to the conclusion that tumors (cancer nodules) are generally larger in size than blood vessels in the lungs and have some semblance of roundness. These properties gave us an idea on how to identify the potential tumor in the CT scan. We set a threshold value for both area and roundness metric and compared the 5 potential cancer nodules against these thresholds, where

Area threshold = 100 pixels

Roundness metric = 0.25

These values are the averages that we have gotten after testing on the cancer CT images. We get the information on each of the 5 objects by first recording their boundary information and applying "`bwboundaries`". After that, "`regionprops`" is used to return the properties, especially the area information on each object. We then loop through each of the object and compute an estimate of their respective perimeters before calculating their roundness value, which has the formula of `4*pi*area/perimeter`^{2}. These obtained results are then compared with the pre-set threshold values. If there is an object which can satisfy all conditions that we have set up, then it is considered as the tumor. In case of a cancer CT scan image, our system will trace the boundary of the object determined by system as cancer nodule on the original CT scan image.

3.5 Graphical User Interface

To better demonstrate our system, we have implemented a simple GUI for user to upload a CT scan image for classification. In our GUI, user can upload an image simply by clicking on the 'Upload' button. The original image, segmented binary image, 5 largest objects with their respective roundness value as well as the final result image will be displayed, with the final verdict classifying whether the image is a normal or cancer CT scan at the bottom. Figure 2 shows the GUI that we have implemented.

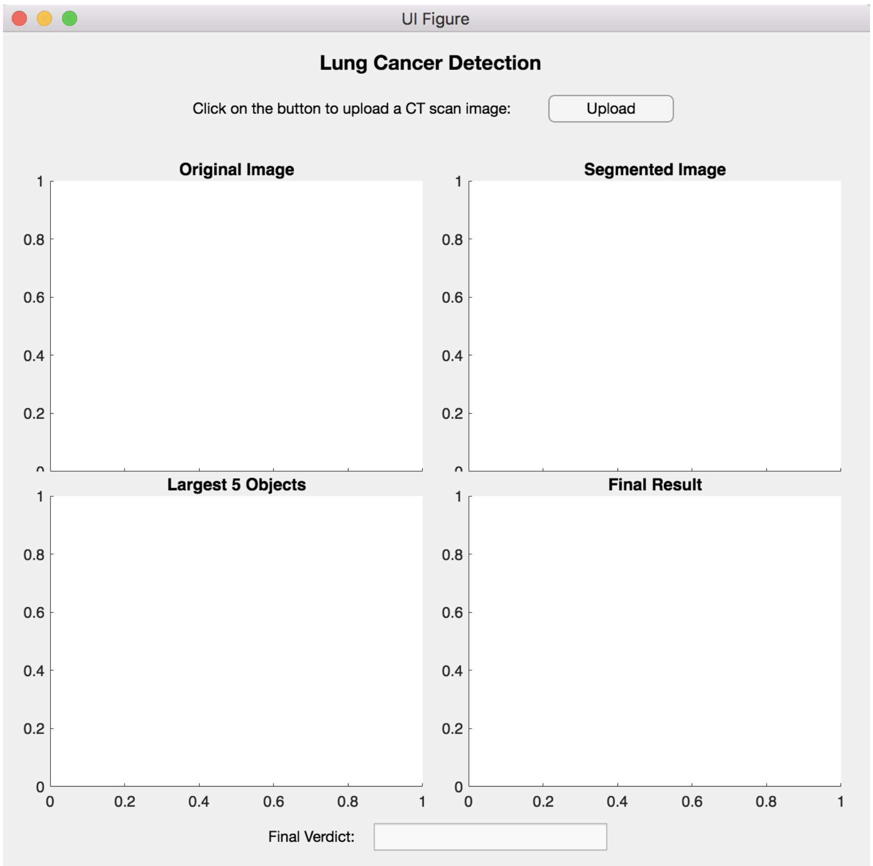

Figure 2

4. Implementation Details

4.1 Pre-processing

4.1.1 Median Filter

The median filter can be implemented in MATLAB using the built-in function "`medfilt2`" which performs median filtering using the median value in m-by-n neighborhood around the specified pixel throughout the corresponding pixel of input image A.

4.1.2 Gabor Filter

The enhancement method using Gabor filter was implemented in MATLAB using the equation(2). All the parameters, are initialized and is then calculated using a nested for-loop function:

```
for x =1:500
      for y=1:500
         x_theta=I_resize(x,y)*cos(theta)+I_resize(x,y)*sin(theta);
        y_theta=-I_resize(x,y)*sin(theta)+I_resize(x,y)*cos(theta);
        gb(x,y)=
      exp((x_theta.^2/2*bw^2+gamma^2*y_theta.^2/2*bw^2))*cos(2*pi/la
      mbda*x_theta+psi);
      end
end
```

4.2 Segmentation

To segment image into objects and background, we use "`imbinarize(I)`" function, where I is a grayscale image and the output is a binary segmented image that utilizes the threshold value found using Otsu's method.

After that, we continue cleaning the segmented image by clearing the unnecessary border using "`imclearborder(BW)`" where BW is a binary image. This function removes anything that is close to or connected to the image border, which is what we are trying to achieve since we are only interested in the objects within the lung regions. We use the default 8-connectivity for this step.

Next step is to identify the largest 5 objects, which are the most likely candidates for cancer nodules. For this, we apply "`bwareafilt(objectsOnly,5)`", where `objectsOnly` is the outcome image of the previous step. We specify 5 in the function, which are exactly the number of objects we are interested in. This function will then output an image containing top 5 objects with largest size.

4.3 Extraction

With the largest 5 objects singled out, we then start our extraction phase by identifying all the boundary information on these objects by applying "`bwboundaries(largest5Objects,'noholes')`", where `largest5Objects` is the image from previous step and the option '`noholes`' specify that we are not interested in the boundaries within the objects itself, if there is any. This function returns two outputs, which we specify as [B, L]. B contains the row and column coordinates of boundary pixels while L is the label matrix of contiguous region.

Using "`regionprops(L,'Area','Centroid')`" function, we are able to get ahold of the area and centroid property values of each labelled region in matrix L. This is to prepare us for the next step in determining whether there exists a cancer nodule within the CT scan.

The system will now go into a loop over all 5 objects and try to compute as estimate of each of their perimeters by using the following two lines:

```
delta_sq = diff(boundary).^2;
perimeter = sum(sqrt(sum(delta_sq,2)));
```

Then, the roundness value of each object is again computed using the formula "`4*pi*area/perimeter`^{2}", in which the area value is the output from `regionprops` function.

We then try to compare the roundness value and area value of these objects against our predefined value to see any of the objects satisfy our conditions, which state that the object has to exceed both of the thresholds. In any case, the roundness value of all 5 objects will be drawn by the side using the following code:

```
text(boundary(1,2)-
35,boundary(1,1)+13,metric_string,'Color','r',...
'FontSize',14,'FontWeight','bold');
```

If there is no object that satisfies the conditions, then system will regard the CT scan as a normal CT scan. Otherwise, system will use the boundary information that we have obtained earlier and trace the outline of said cancer nodule and declare that it belongs to a cancer CT scan image.

5. Analysis of the Outcome with Test Results

To validate the proposed methodology, 69 lung CT scan images were extracted from LIDC-IDRI dataset; 39 images with lung cancer and 30 images with normal lung.

A. *Pre-processing*

All the selected images contains high resolution with no noise. Therefore, for this test, the pre-processing stage for both image smoothing and enhancement were omitted as it did not make significant improvements in the image; neither did it reduce distortions nor enhance the nodules (only the blood vessels as shown in Figure 3).

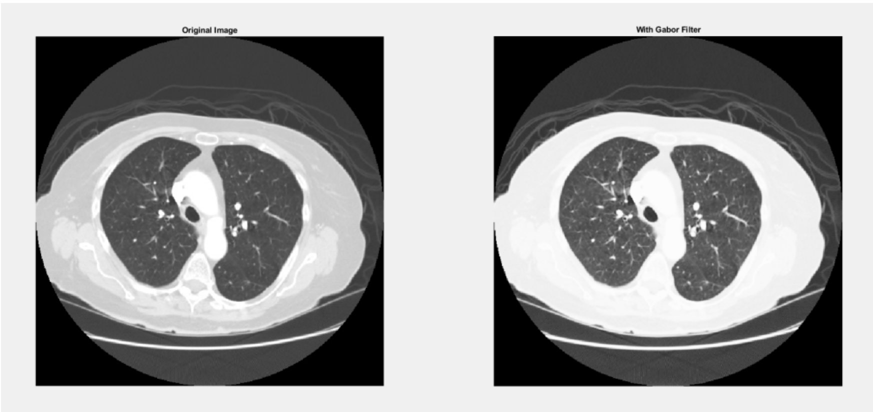

Figure 3

B. *Segmentation*

The segmentation details are shown in Figure 4. The original grayscale image is converted into a binary image using Otsu's method using its threshold value. Then the required feature is extracted, the lung-only image. Finally, using segmentation, 5 largest objects is extracted.

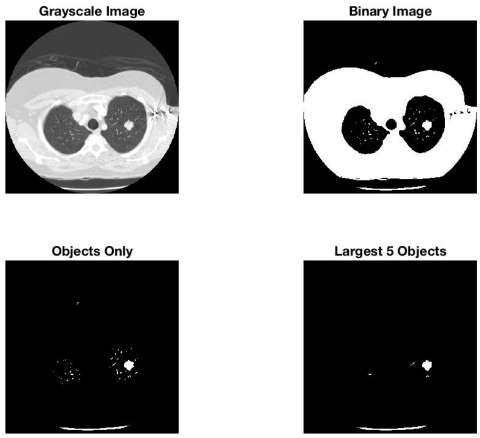

Figure 4

C. Extraction

The extraction details of a cancer detected image are shown in Figure 5 where the 5 identified objects of the segmented binary image is used to compare its area and roundness values with the threshold value to identify whether the object is an abnormal or normal lung. If cancer is detected, a red boundary is drawn around the object.

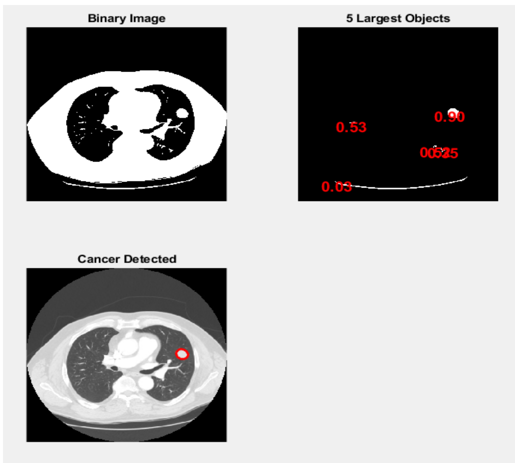

Figure 5

The successful case using the GUI is shown in Figure 6. Selected image is uploaded onto the system, then it analyses using each processing stage to determine whether the CT scan image is abnormal or normal.

Figure 6

Table 1 shows the rate of measurement of the test tested CT scan images based on true acceptance rate (TAR), true reject rate (TRR), false acceptance rate (FAR), false reject rate (FRR).

Rate of Measurement	TAR	TRR	FAR	FRR
Number of Images	32	24	6	7

Table 1

Using the rate of measurement, the classification accuracy can be computed within the tested dataset:

Classification Accuracy = [(TAR + TRR)/(total number of images)] * 100%

= 56/69*100%
= 81.2%

6. Limitation

Pre-processing

CT images contain various types of noises, such as salt and pepper, Gaussian and speckle noise, mainly due to the random noise caused by the statistical inaccuracies and limitation during calculation of finite number of transmitted x-ray quanta (Fong Y. Tsai, 1981). For each different type of noise, it has a specific filter that can perform better. Hence, using median filter may reduce noise in salt and pepper noise[6]; however, Wiener filter may be a better alternative for reduction in Gaussian noise [14]. Therefore, our system may not be able to clean up all the noises that are present in CT images that are of lower quality.

Thresholding

One of the major limitations for this approach is that it only works best with images that have bimodal histogram, ie. those with two distinct regions that can clearly separate the pixels into either background or object pixels. For those images that do not possess this quality, the threshold value obtained will not be able to segment the images in a precise manner, thus reducing the accuracy. Other than that, Thresholding also assumes that input images are of uniform illumination since it largely depends on the pixel intensity. This is not a serious issue in our case since we are using CT scan images so the illumination is almost guaranteed to be uniform.

Extraction

Counting number of black pixels (original proposed method):

In our proposal, we were trying to come up with an average value that would be the classifying factor to determine the input CT scan as one of the two classes. The idea behind this classifying technique is that a value is set as a kind of threshold. CT scan image is segmented until only both lungs region is retained. The number of black pixels within the lungs has to exceed the value set to be considered as normal CT scan. The theory is that because blood vessels and cancer nodules are regarded as objects within the lungs, so if there exists a tumor (which generally is bigger in size), then the number of black pixels in the lung regions should be smaller than the present value. Essentially, this method works best when the size of the lung regions in dataset is relatively constant. For our case, the tumor in each cancer CT scan is located at different locations, thus captured in slices that contain varying sizes of lung region, which resulted in us discarding this method since it no longer works for our dataset Figure 7 shown to illustrate our point.

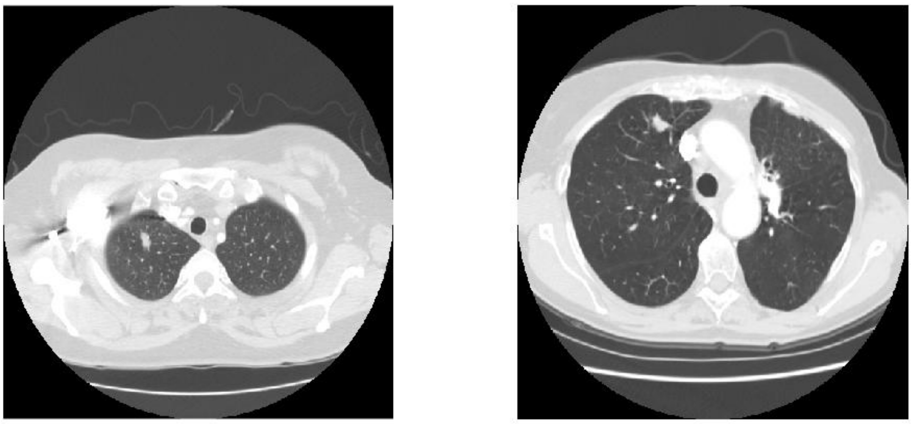

Figure 7

Roundness metric:

Since the area and roundness threshold values are not obtained from a larger set of data, the values are not as accurate as we would like them to be. Therefore, these values do not apply to some CT scan image, which leads to the case of false results. An example can be seen from Figure 8, where the roundness value for the tumor (0.15) is lesser than the threshold, thus being falsely classified as a normal CT scan image.

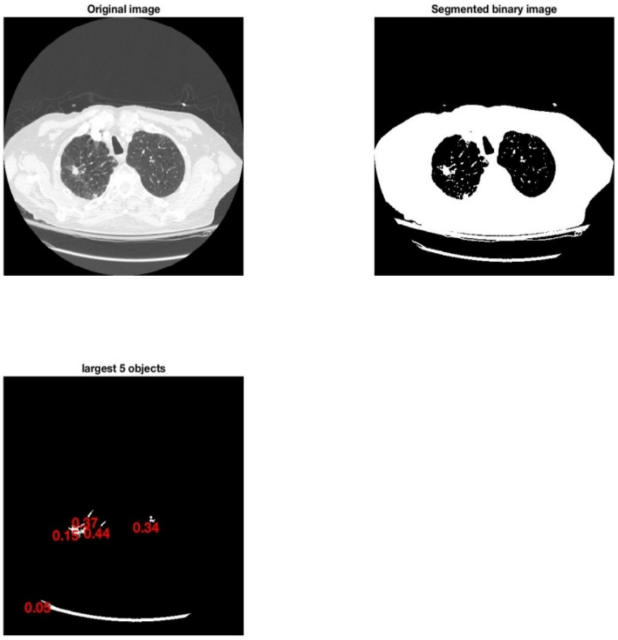

Figure 8

Another example would be Figure 9 where the blood vessels in the input CT scan slice has a roundness value greater than the threshold and once again, resulted in inaccurate classification.

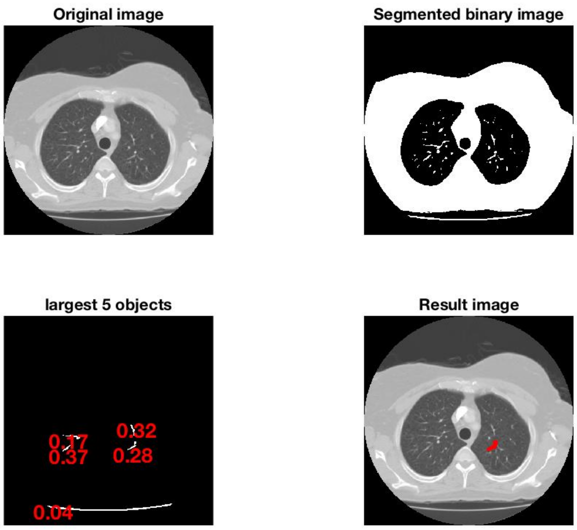

Figure 9

Position of tumor:

We discovered that our system fails to detect cancer nodules which are situated close to the border. This is because in our implementation, we remove the border as we are only interested in the objects within the lung regions. However, because of its position, after segmentation, the tumor appears to be connected to the border. Therefore, function "`imclearborder`" regards it as a border element and removes it as well, thus leading to another false classification. An example can be seen from Figure 10.

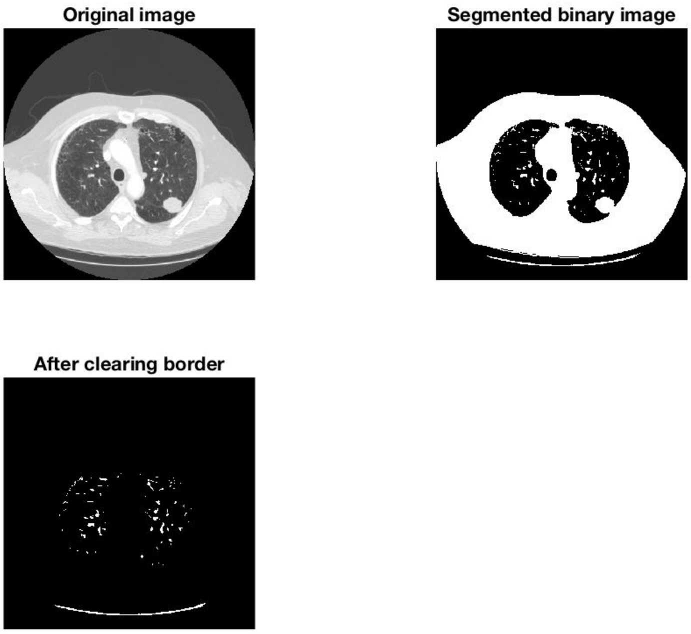

Figure 10

7. Improvement

To the limitations stated in the extraction phase, we believe that a larger set of data can be used to rectify the situation. We can also counter the problems by allowing all slices of CT scan as input instead of just one slice. This is because tumor usually appears out of nowhere, contrary to the blood vessels within the lung regions which appears and fades away with each increasing slice. We can also train a neural network to help detect the changes in the slices for better extraction in the future.

Another improvement to our system will be to identify the stages of cancer for cancer CT scan. We can use the area, perimeter and eccentricity information given by professional to achieve this goal. This refinement will undoubtedly be able to help more patients receive earlier detection for better recovery rate.

As mentioned in the limitation, the position of the tumor near the border has failed to detect the cancer lung nodule. This can be improved by applying marker controlled watershed segmentation algorithm as is separates the overlapping object of the image.

8. Conclusion

Attention is highly given for lung cancer detection in an attempt to increase recovery rate using image improvement technique on CT scan images. The proposed methodology of lung cancer detection, using a new classification method, was able to detect cancer nodules. The CT scan images is pre-processed using median filter and Gabor filter. Then the pre-processed image goes through segmentation stages using Thresholding. Finally, the segmented images is extracted to classify its features to find whether the potential nodules are indeed cancer tumour. The classification accuracy is 81.2%. For future work, this technique can be further improved by implementing different cancer stages. Also, increasing dataset will improve its classification accuracy.

Reference

1. Society, A.C. *Lung Cancer*. n.d; Available from: https://www.cancer.org/cancer/lung-cancer.html.
2. American Cancer Society. *Key Statistics for Lung Cancer*. 2018 January 4, 2018; Available from: https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html.
3. Pu, J., et al., *Adaptive border marching algorithm: automatic lung segmentation on chest CT images.* Computerized Medical Imaging and Graphics, 2008. **32**(6): p. 452-462.
4. Acess, T.C.I.A.T.P. *LIDC-IDRI*. May 22, 2018.
5. Mahersia, H., M. Zaroug, and L. Gabralla, *Lung cancer detection on CT scan images: a review on the analysis techniques.* (IJARAI) International Journal of Advanced Research in Artificial Intelligence, 2015. **4**(4).
6. Makaju, S., et al., *Lung Cancer Detection using CT Scan Images.* Procedia Computer Science, 2018. **125**: p. 107-114.
7. Avinash, S., K. Manjunath, and S. Senthilkumar. *Analysis and comparison of image enhancement techniques for the prediction of lung cancer*. in *Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017 2nd IEEE International Conference on*. 2017. IEEE.
8. Chaudhary, A. and S.S. Singh. *Lung cancer detection on CT images by using image processing*. in *Computing Sciences (ICCS), 2012 International Conference on*. 2012. IEEE.
9. Kulkarni, A. and A. Panditrao. *Classification of lung cancer stages on CT scan images using image processing*. in *Advanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference on*. 2014. IEEE.
10. Gajdhane, M.V.A. and L. Deshpande, *Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques.* IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN, 2014: p. 2278-0661.
11. Al-tarawneh, M.S., *Lung cancer detection using image processing techniques.* Leonardo Electronic Journal of Practices and Technologies, 2012. **11**(21): p. 147-58.
12. Miah, M.B.A. and M.A. Yousuf. *Detection of lung cancer from CT image using image processing and neural network*. in *Electrical Engineering and Information Communication Technology (ICEEICT), 2015 International Conference on*. 2015. IEEE.
13. Abdillah, B., A. Bustamam, and D. Sarwinda. *Image processing based detection of lung cancer on CT scan images*. in *Journal of Physics: Conference Series*. 2017. IOP Publishing.
14. Chahar, P.S. and V. Thakare, *Performance Comparison of Various Filters for Removing Gaussian and Poisson Noises.* International Research Journal of Engineering and Technology, 2015. **2**(5): p. 1101-1105.