# Segmentation and Image Analysis of Pneumonia

Jiujiu Pan, Tianchang Li, Maggie Guo
BMI/CS 567 Final Project

***Abstract*—The computer-based identification of the boundaries of the lung from surrounding thoracic tissue on x-ray images, which is called segmentation, is a vital first step in radiologic pulmonary image analysis. In order to apply image processing methods we learned in BMI/CS 567, we utilized thresholding-based and region-based segmentation methods. However, the lungs with moderate to high abnormalities (opacification) would blur the boundaries, making it specifically hard to segment clearly. Therefore, we included a self-defined, carefully calibrated ratio vector methods to cope with such potential problems, with extra emphasis on the accuracy of the methods in cases with both control and abnormality groups.**

## I. Introduction

The computer-based identification of the boundaries of the lung from surrounding thoracic tissue on x-ray images, which is called segmentation, is a vital first step in radiologic pulmonary image analysis. Nearly all CT images are now digital, thus allowing increasingly sophisticated image reconstruction techniques as well as image analysis methods. In the step of segmentation, the lung is detected, and its anatomic boundaries are delineated, either automatically or manually.

The purpose of this project is to employ image analysis methods to identify lung abnormalities from chest X-ray images. To make the analysis more efficient, we planned to incorporate a thresholding-based method that could automatically identify the edge of lungs[2]. Errors in lung segmentation would generate false information with regard to subsequent identification of diseased areas, but such error is unfortunately very common since the infected lungs would result in an X-ray image of blurred edges due to opacification of tissues (See more description of images and dataset in Section II).

Due to the limitation of available methods we could utilize and studied in BMI 567[1], we shared the threshold features between the normal and infected lungs, which means that the infected lungs might only have the clear part segmented. We then chose to cope with potential segmentation errors by estimating the original lung area from boundaries of segmented figures and calculated the ratio of segmented area to the estimated true lung area. Both normal and pneumonia groups would have their results represented by ratio. Accuracy and statistical tests would be performed on these methods. We evaluate that this estimation method without too intricate algorithms has higher efficiency and could potentially serve as a preliminary insight to a patient's pulmonary image analysis.

## II. Data

The dataset of this project is adapted from [3] Illustrative Examples of Chest X-Rays in Patients with Pneumonia, Related to Figure 1. The normal chest X-ray (left panel) depicts clear lungs without any areas of abnormalities in the image. Pneumonia (right) manifests with moderate opacification in both lungs. The dataset is organized into 2 folders (test and valuation) and contains subfolders for each image category (Pneumonia/Normal). There are 640 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal).

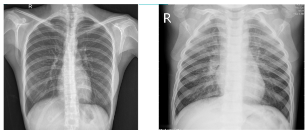

Fig. 1: Normal lung (left) and infected lung (right)

## III. Methods

1. Image Resize and Adjust

Because of the image in the data sets have various dimensions, and in order to apply universal threshold and region growing method for segment the lungs correctly, we first uses imresize() and imadjust() built in function in MATLAB to normalize all images to the same dimension and better Luminance.

2. Thresholding

After normalization, we use imhis() to read some peak frequency and pass the histgram output value to otsuthresh() function and generate the binarized image, which separate the lung part with the human body.

3. Segmentation

We used region growing technique to segment the two lobes of the lungs. This method creates a 3x3 mask starting from a manually selected pixel within each lung and grows masks with the same size centered at each pixel in the current masks with brightness of 1. Then we use imfill() to fill out the holde inside the lungs that have not been grown successfully. Ideally it should grab the whole lung area from the thresholded images

4. Calculate the ratio

After getting the final extracted lung, we going a simple calculation by dividing the are of the lung to the area that the boundary of the lung spans. As expected the normal lung would have higher and more stable ratio because of no infected area. This will generate a set of ratios for all images for doing z-test later

5. Predict

We used population z-test as the classifier in our project. Our null hypothesis was that each image did not belong to the normal lung population.

$$H_0 : x_i = \mu . H_A : x_i \neq \mu$$

The z score of each test sample was calculated in the following manner:

$$z_i = \frac{x_i - \mu}{\sigma * \sqrt{n}}$$

where $z_i$ is the z score for i th sample, $x_i$ is the ratio of normal lung computed for the i th sample, $\mu$ is the population mean of the ratio, $\sigma$ is the population standard deviation of the ratio, and $n$ is the test sample size. We calculated the mean and standard deviation from 50 images of normal lungs and, based on the central limit theorem, assumed them to be unbiased estimators of the population distribution parameters. Since we tested each sample individually, $n$ is always one. The computed z score will be compared with normal distribution. We rejected the ones that were shown to be different from normal distribution with 95% confidence.

We will conclude our results with the accuracy our prediction achieved in both normal and pneumonia lungs. The false positive rate and false negative rate of this z-test classifier will be computed.

## IV. Experiments

Firstly, we processed the images labeled with normal lungs and computed the ratio of lung area.The original images were taken with various background and contrast. To best scale the brightness and prepare for thresholding, we applied imadjust() function with low-in = 0.4, high-in = 0.8, and gamma = 0.9. The lungs in our images are mostly brighter than the background but darker than the bone structure. These parameters are the ones we found optimal for filtering out pixels too dark or too bright. Gamma = 0.9 brought images a little brighter than original ones as a whole.

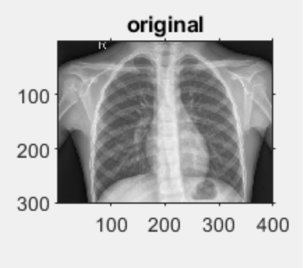

Fig. 2: original image

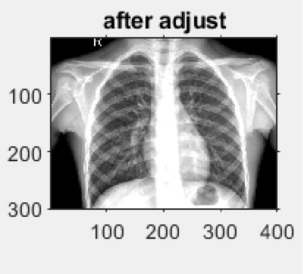

Fig. 3: enhanced contrast image

Secondly, we uses the imhist() to retrieve the top 100 frequency in the image and computes a global threshold T from histogram counts using otsuthresh() .Otsu's method chooses a threshold that minimizes the intraclass variance of the thresholded black and white pixels. Then we use imbinarize() to convert the grayscale image to a binary image which white part is the body part except for lung.

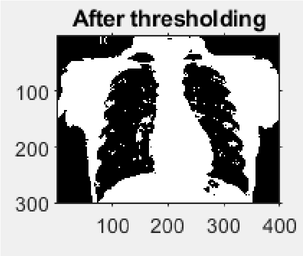

Fig. 4: thresholded image

Most normal lungs are easy to segment from background by thresholding. With the lung area split up with the rest of the image, we applied region growth starting from one pixel within each lobe of the lung. These two pixels were manually selected that worked for all images. The result is shown below.

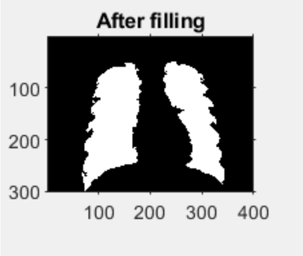

Fig. 5: final processed image

However, there are some cases that the image is too whiteish because of the extreme seriousness of the Pneumonia, during which cases our code cannot segment the blurry lung well and will generate smaller than expected lung area. But this will not affect our prediction process because those data will have small ratio which we deem as Pneumonia cases. The example image is shown below.

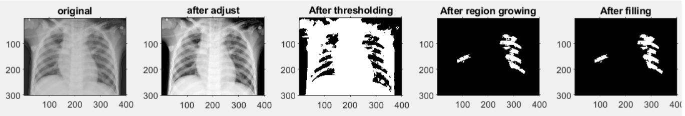

Fig. 6: image analysis with Pneumonia

With all the normal lungs segmented, we calculated the ratio of the lung area to the rectangular area that the boundary of the lung spans in each image. Since people with Pneumonia have smaller lung areas reflected on their X-ray images, this ratio then was used as the primary estimator or how good or bad a lung is.

The mean and standard deviation of these 50 ratios were then taken as the unbiased estimators of the population distribution of normal lung images. We tested 30 normal images and 30 pneumonia images with this specified distribution respectively. With the mean and standard deviation we computed, all pneumonia images are rejected, meaning classified as abnormal, but 29 out of 30 normal lungs were rejected too. This might result from the high variance of the testing images since we were only testing one each time instead of a batch. We then loosed the input standard deviation a little by multiplying it with two times square root of training size:

$$z_i = \frac{x_i - \mu}{\sigma * 2 * \sqrt{N}}$$

where N is the sample size of training set (normal set). This gave us the final optimal results with an accuracy of 63.3% in normal lungs and 73.3% in pneumonia lungs, which suggests a false positive rate of 26.7% and a false negative rate of 36.7%.

## V. Conclusions

As seen by the results and accuracy test, our devised methods could distinguish between normal and infects lungs at an acceptable level (˜70% accuracy). Such ratio-based estimation method could potentially serve as a preliminary assessment. Nevertheless, one should be aware of the limitations of its precision on extrapolating the original lung area. Therefore, should the method be put into further usage, a more rigorous extrapolation algorithm for estimating original lung area is required in the future.

## References

[1] Daniel Pimentel-Alarcón Teaching Materials. (2020). Retrieved from https://danielpimentel.github.io/teaching.html

[2] Global histogram threshold using Otsu's method. (n.d.). Retrieved from https://www.mathworks.com/help/images/ref/otsuthresh.html

[3] Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification, *Mendeley Data, v2*, https://data.mendeley.com/datasets/rscbjbr9sj/2