Clustering Techniques for COVID-19 CT scan analysis

Arman Mohammadigilani, s298786,
ICT for Health

December 22nd, 2021

1 Introduction

With the increasing prevalence of coronavirus disease-19 (COVID-19) infection worldwide, early detection has become crucial to ensure rapid prevention and timely treatment. However, due to the unknown gene sequence of the supposed coronavirus, the reference standard test has not been established for diagnosis. Several studies have suggested pneumonia as the underlying mechanism of lung injury in patients with COVID-19 Accordingly, it is believed that the pulmonary lesions caused by COVID-19 infection are similar to those of pneumonia. More than 75% of suspected patients showed bilateral pneumonia. In this context, the promising findings of several studies have highlighted the growing role of chest computed tomography (CT) scan for identifying suspected or confirmed cases of COVID-19 infection.

The common typical chest CT scan findings are summarized as: Peripheral distribution, Bilateral lung involvement, Multifocal involvement, Ground glass opacification-GGO (instead of appearing uniformly dark), Crazy paving appearance (appearance of ground-glass opacity with superimposed interlobular septal thickening and intralobular septal thickening), Interlobular septal thickening(numerous clearly visible septal lines usually indicates the presence of some interstitial abnormality), Bronchiolectasis (dilatation of the usually terminal bronchioles (as from chronic bronchial infection)). In other words, lung alveoli are partially filled with exudate or they are partially collapsed and the tissue around alveoli is thickened.

Not all the patients affected by COVID-19 show interstitial pneumonia, but its presence is a fast way to diagnose COVID-19. Nasopharyngeal swab analysis requires some hours in the lab plus the time to deliver the swab to the lab; on the contrary, any hospital has CT scanners and the radiologist can immediately detect the presence of ground glass opacities. However, it would be useful to design an algorithm to help radiologists in this task. In the next sections a method is described that identifies these opacities for the subsequent analysis by the radiologist. The software was developed in Python, using the Scikit-learn library.

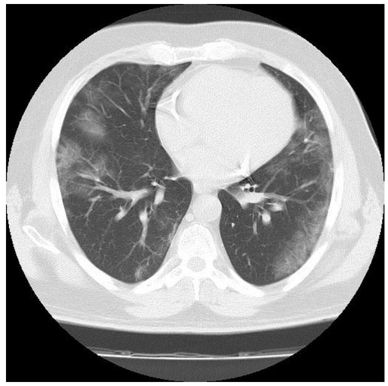

Figure 1: Example of ground glass opacity (light grey opaque areas in the lungs).

2 Method

An example of ground glass opacity can be seen in the CT scan of Fig. 1. Indeed, a CT scan is made of many slices of the patient chest in the axial plane, and Fig. 1 is just one of these slices. Specific COVID-19 CT scans were downloaded from [1]; for each patient around 300 slices are present, each one being a grey scale image with 512×512 pixels.

The proposed method is made of two main steps:

1. identify the position of lungs (image segmentation)

2. find the greyish areas in the figure portion corresponding to lungs

and both tasks are solved using two clustering algorithms, namely K-means [2, Chapter 11] and DBSCAN (Density-based spatial clustering of applications with noise) [3].

2.1 Identify lungs

The first step to automatically find the position of lungs in the image is to quantize its colors using K-means with 5 clusters: the resulting image (Fig. 2) is very similar to the original one, but it is made of just 5 colors, the darkest being the background. Lungs include dark grey pixels that do not appear elsewhere and therefore the K-means cluster with the second darkest color at least partially corresponds to lungs, as shown in Fig. 3 (purple in the image corresponds to 1 in a 512×512 matrix).

Application of DBSCAN on the coordinates of purple pixels in Fig. 3 (neighborhood radius $\epsilon = 2$, minimum number of points 5) allows to separate the borders of the bed and chest from the lungs, which are the two most populated clusters. Actually, not all the purple points of a lung are given to the same cluster by DBSCAN, but the position of at least a portion of the two lungs can be identified (see Fig. 4). If DBSCAN is now applied to the coordinates of pixels with either the darkest or the second darkest quantized colors, many clusters are generated, but lungs are those clusters whose centroid (barycenter) is closer to the centroid of the two lung portions in Fig. 4. The obtained image is shown in Fig.

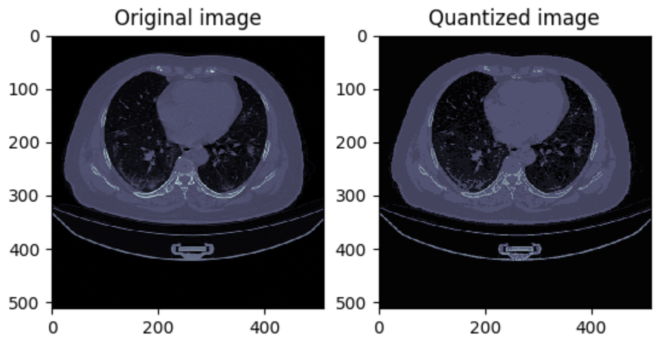

Figure 2: Original (left) and color quantized (right) images.

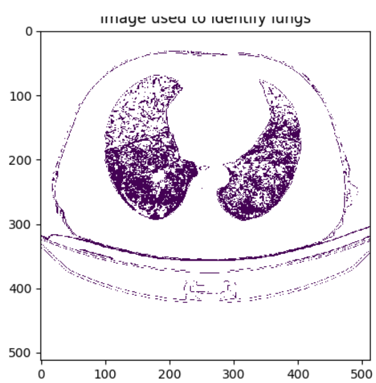

Figure 3: Region with the second darkest color after quantization through K-means.

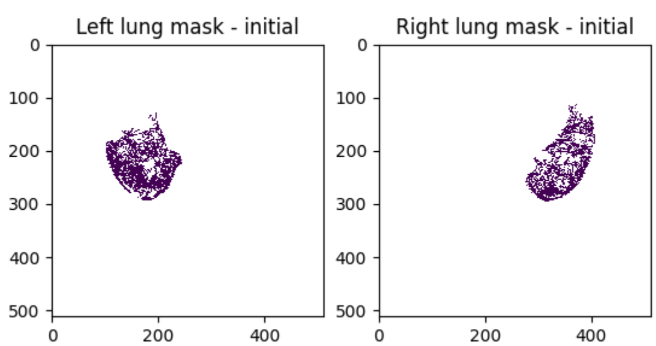

Figure 4: Initial identification of lungs.

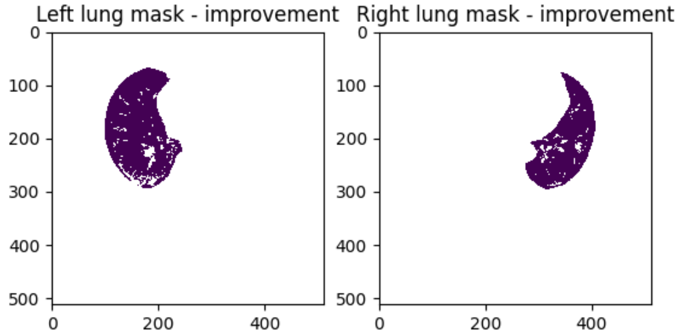

Figure 5: Intermediate identification of lungs.

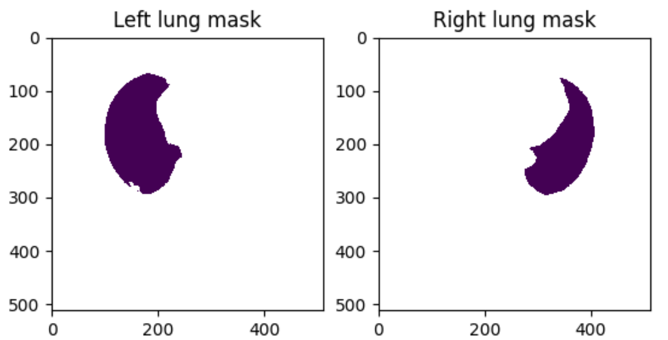

Figure 6: Final identification of lungs.

5, which is almost correct, apart from the presence of "holes" inside the lungs, where the original image has light grey colors.

Application of DBSCAN on the coordinates of pixels that are NOT purple in Fig. 5 allows to solve the problem: the algorithm finds a big cluster that surrounds each lung and many small clusters (maybe classified as noise) inside the lungs. Then the lung mask is the set of pixels that are NOT included in the most populated cluster found by DBSCAN. This final result is shown in Fig. 6. Note that one undesired notch is present in the lower left part of the lung on the left and another small one appears in the lung on the right; these imperfections are due to almost white colors in these pixels in the original image.

2.2 Find the ground glass opacities

The true colors of the CT scan in the lung masks are shown in Fig. 6, whereas 'viridis' colormap was used to generate the image in Fig. 8. In this second figure the opacities are more clearly visible and this suggests that it is sufficient to choose the correct range of values in the grey scale to identify them.

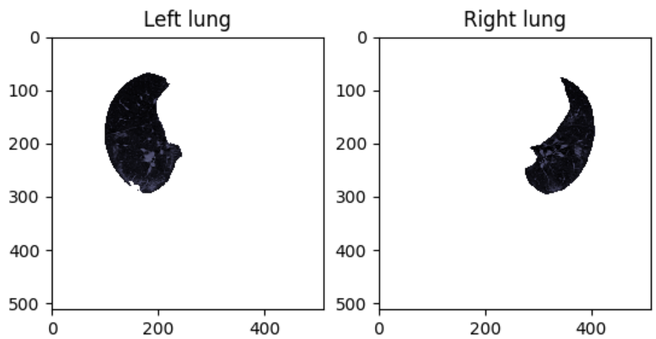

Figure 7: Image of lungs with bone colormap.

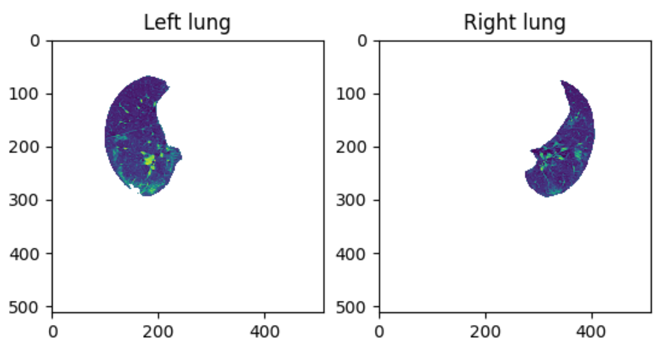

Figure 8: Image of lungs with viridis colormap.

In particular, we chose the range $[-750, -400]$ to detect pixels corresponding to possible infection area; then filtering was used with a kernel with size 2 and the final infection mask (see Fig. 9) was obtained as the set of pixels with values higher than a threshold set equal to 0.25. Fig. 10 shows the original image and the superimposed infection mask.

By analyzing figures 9 and 10 with different values of parameters, optimal values were chosen. We should choose a reasonable color range to recognize infection pixels. In addition, some points are scattered which should be removed by using the image filter. If we choose a low kernel parameter, scattered points wouldn't be removed and a high square kernel value results in removing some essential points.

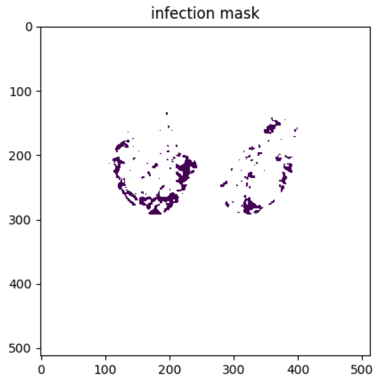

Figure 9: Infection mask.

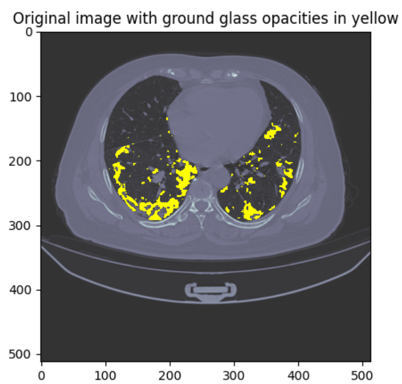

Figure 10: Infection region (in yellow) superimposed to the original image.

2.3 An index to define the severity of pneumonia

The Percentage coefficient is designed to show infection severity. It is the relation of the infected part of the lung to the whole lung so it is calculated by the number of all infected pixels divided by the number of all pixels of the lung image.And it is clear that an increase in this coefficient indicates an increase in infection in the patient's lungs.

$$\text{right lung infection index} = \frac{right\ lung\ infected\ pixels}{right\ lung\ totall\ pixels} = \frac{1522}{13694} \times 100 = 11.11\%$$

$$\text{left lung infection index} = \frac{left\ lung\ infected\ pixels}{left\ lung\ totall\ pixels} = \frac{2894}{19793} \times 100 = 14.62\%$$

3 Conclusions

In conclusion, we realized that choosing reasonable parameters play a very important role to have an optimal figure of the infected lung. In addition, it is essential that these parameters are selected by the opinion of radiologists and close to the figure they desire. In this case, They can more easily detect areas of lung infection. It also allows quick treatment for patients in the most efficient and appropriate way.

According to the infection severity coefficient, 11.11% of right lung is infected while this value is 14.62% for the left lung.

References

[1] `https://www.kaggle.com/andrewmvd/covid19-ct-scans`

[2] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012

[3] Martin Ester , Hans-Peter Kriegel , Jörg Sander , Xiaowei Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996