Report 3: Clustering techniques for COVID-19 CT scan analysis

Seyed Mohammad Mehdi Hosseini, s301769,
ICT for Health attended in A.Y. 2021/22

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.

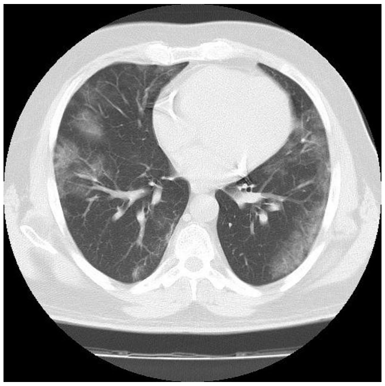

Figure 1: Example of ground glass opacity (light grey opaque areas in the lungs).

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.

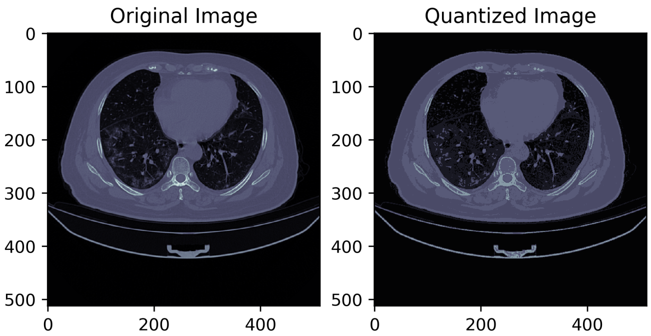

Figure 2: Original (left) and color quantized (right) images.

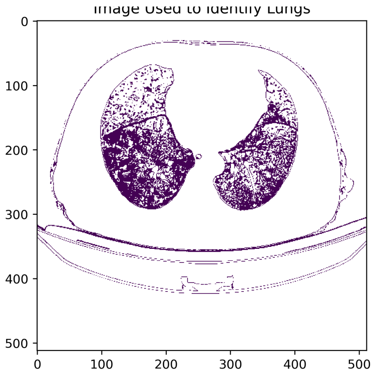

Figure 3: Region with the second darkest color after quantization through K-means.

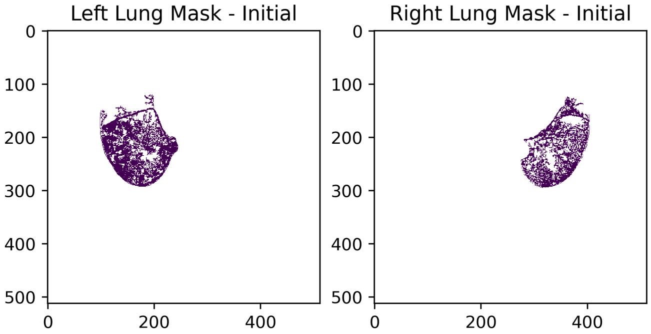

Figure 4: Initial identification of lungs.

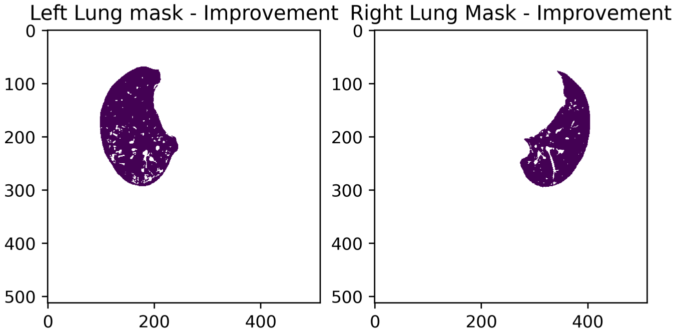

Figure 5: Intermediate identification of lungs.

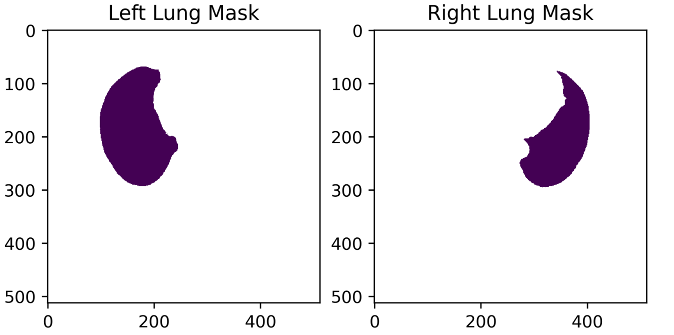

Figure 6: Final identification of lungs.

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.

In particular, we chose the range $[-700, -350]$ to detect pixels corresponding to possible

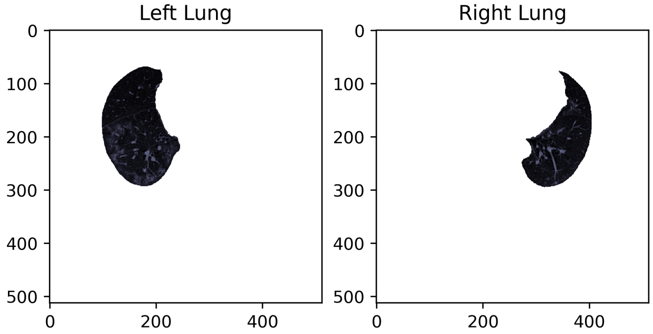

Figure 7: Image of lungs with bone colormap.

infection area; then filtering was used with a kernel with size 20 and the final infection mask (see Fig. 9) was obtained as the set of pixels with values higher than a threshold set equal to 0.11. Fig. 10 shows the original image, the superimposed infection mask and also the actual infection mask.

To evaluate the quality of the segmentation, several metrics have been included. The first one is the Dice Similarity Coefficient (DSC) Score [4] which is one of the most famous and state-of-the-art metrics to evaluate the segmented images. The DSC ranges from 0, where no spatial overlap between two binary images exists to 1, where both images are overlapped completely. Comparing both ground truth and the image with ground-glass opacities, the algorithm results in a 0.84 score which is a reasonable score considering the algorithm complexity. On top of that, two pixel-wise metrics, sensitivity and specificity have been included in this research. The presented method has been achieved 0.81 sensitivity and respectively 0.99 specificity in total pixels for this specific slice. It should be noted that different parameters result in different outcomes e.g. smaller kernels result in scattered ground-glass opacities and consequently, the larger result in more consistent segments. The square kernels between 15 to 25 could obtain better results in the bargain. The opacity threshold is also another important to declare opacity in an image. Considering the 20 kernel size, thresholds between 9.5 to 10.5 could obtain interesting results with different DSC and sensitivity scores. Just as importantly, the algorithm has failed to segment the little part of infections, especially in the right lung but it could find the most infected parts of the lungs which are mostly located in the left lung. The corresponding founded infections could be distinguished with orange color, and the actual segments could be observed with dark green color (See 10). The light green color is where the algorithm could successfully segment the infections in both lungs.

2.3 Lungs involvement

About January 2020, the first findings of the COVID-19 were officially published and the majority of hospitalized patients had bilateral lung involvements and ground-glass opacities on their chest CT scans [5]. Accordingly, most of the clinicians used the CT images as a

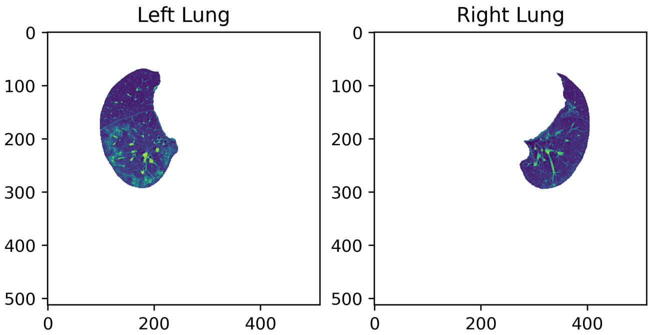

Figure 8: Image of lungs with viridis colormap.

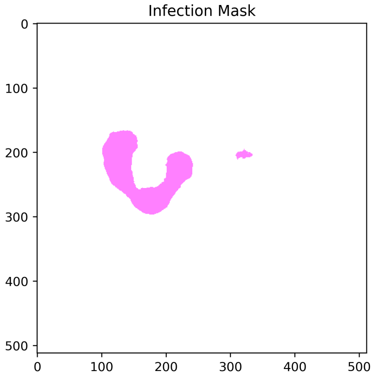

Figure 9: Infection mask.

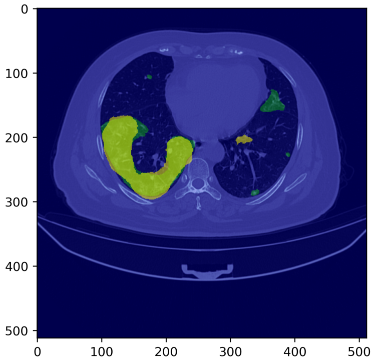

Figure 10: Infection region (in yellow) superimposed to the original image and actual infection mask (in green)

major reference to find the stage of the COVID-19 disease in patients by considering the lungs involvement percentage. In the proposed model, the total number of ground-glass opacities and lungs pixels found in the CT scan image will be taken into the account. Dividing these two values, the lungs involvement percentage will be obtained and it could determine how good or bad is the patient's condition. Additionally, the model could find which lung is more involved with the infections. For this particular slice (See 10), about 14% of the total lung have been infected, and the left one is more infected than the right one. The model also has been equipped with a simple human-speech aided system which could automatically help the operator get the summary of results.

3 Conclusions

Most of the people infected with COVID-19 will experience a mild to moderate illness and recover without any special clinical treatments but on the other hand, there are a group of people who could become seriously ill and will need proper medical treatments under the supervision of treatment teams and required facilities. Today, by emerging new and contiguous variants of COVID-19 just like Omicron (B.1.1.529): SARS-CoV-2, there is a huge concern regarding the speed of COVID-19 diagnosis to prevent its spreading and also the terrible lungs involvements between the patients. Scientists have found CT scans as one of the most important references to recognize COVID-19 patients and their stage of the disease. In this research, two clustering algorithms, Kmeans and DBSCAN have been applied to find the position of lungs and also the ground-glass opacities. Considering the complexity of the above-mentioned algorithms which is much more simple than U-Net in image segmentation tasks, the results are promising and reasonable. There are several metrics to evaluate the model like Dice Similarity Coefficient Score (DSC), Sensitivity, Specificity and etc. which could demonstrate how much the model is good at finding the interstitial abnormalities.

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

[4] Zou, Kelly H., et al. "Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index1." . *Academic Radiology, no. 2, Elsevier BV, Feb. 2004, pp. 178–89. Crossref, doi:10.1016/s1076-6332(03)00671-8*

[5] Kwee, Thomas C., and Robert M. Kwee. “Chest CT in COVID-19: What the Radiologist Needs to Know.” *RadioGraphics, no. 7, Radiological Society of North America (RSNA), Nov. 2020, pp. 1848–65. Crossref, doi:10.1148/rg.2020200159.*