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

Francesco Donato, s304810,
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.

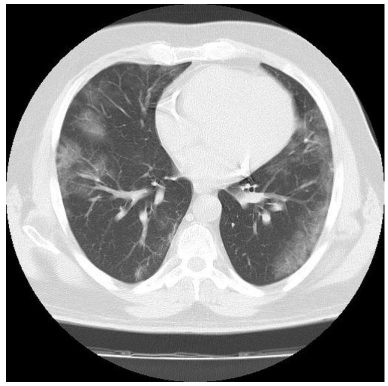

Figure 1: Example of ground glass opacity (light grey opaque areas in the lungs).

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.

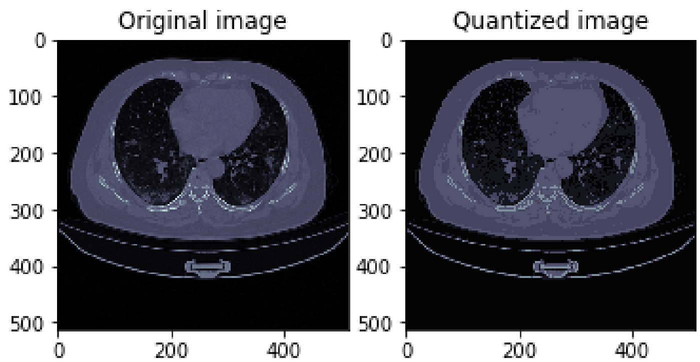

Figure 2: Original (left) and color quantized (right) images.

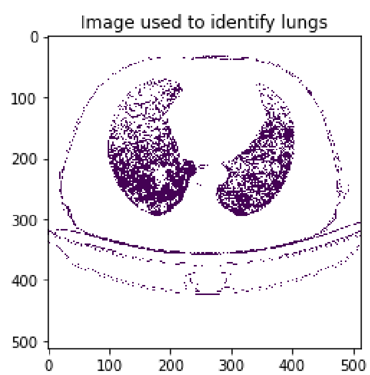

Figure 3: Region with the second darkest color after quantization through K-means.

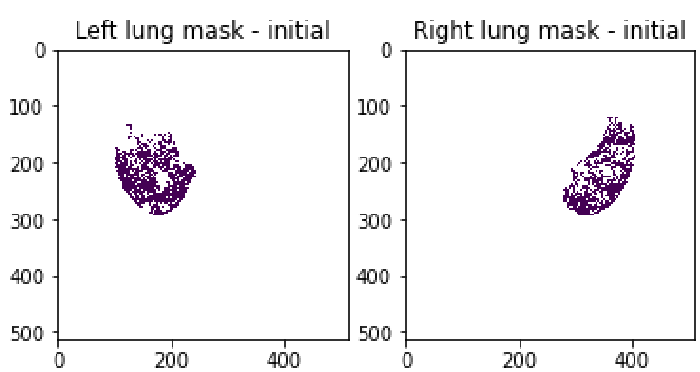

Figure 4: Initial identification of lungs.

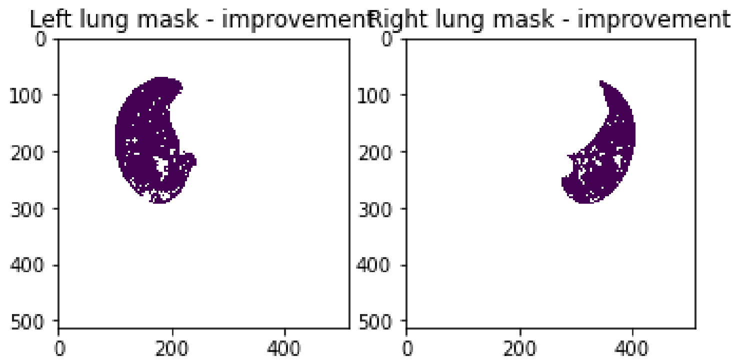

Figure 5: Intermediate identification of lungs.

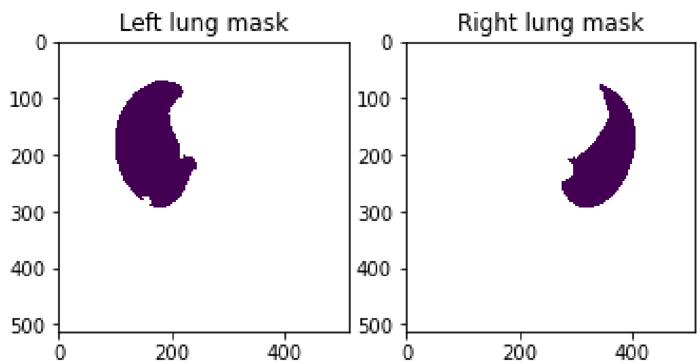

Figure 6: Final identification of lungs.

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. 7, 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 $[-700, -250]$ to detect pixels corresponding to possible

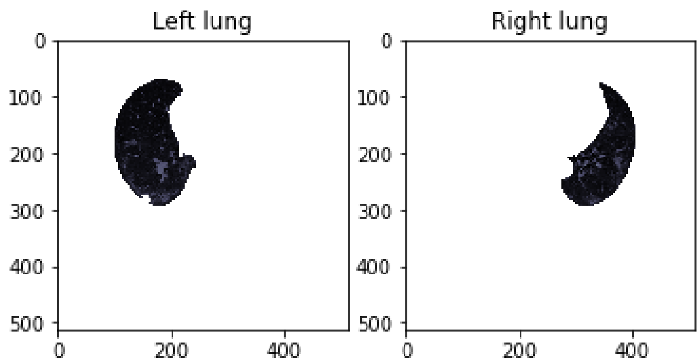

Figure 7: Image of lungs with bone colormap.

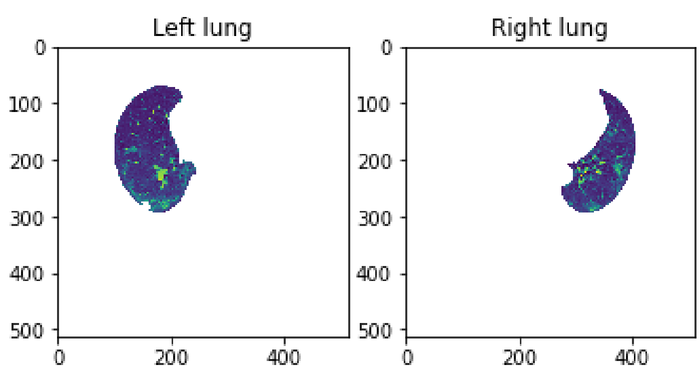

Figure 8: Image of lungs with viridis colormap.

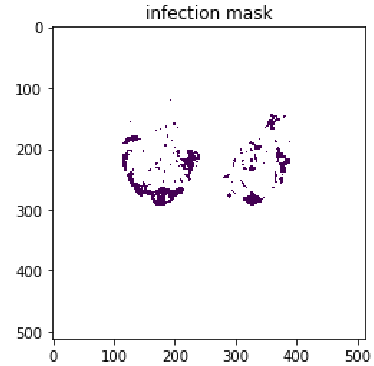

Figure 9: Infection mask.

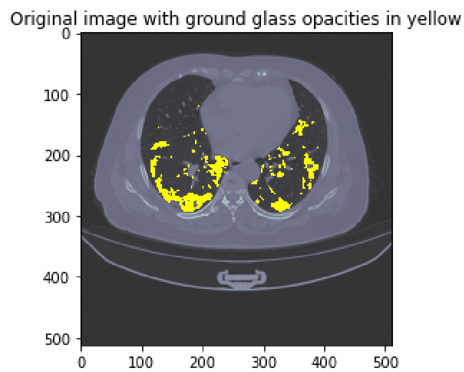

Figure 10: Infection region (in yellow) superimposed to the original image.

infection area; then filtering was used with a kernel with size 5x5 (RowsxColumns) and the final infection mask (see Fig. 9) was obtained as the set of pixels with values higher than a threshold set equal to 0.2. Fig. 10 shows the original image and the superimposed infection mask.

It is possible to see how the infection mask obtained in Fig. 8 is too scattered. For this reason, the image was filtered, obtaining the infection mask shown in Fig. 9, which is less smooth than that given in the dataset, but it might be more realistic, in the end. This infection mask was later used to highlight (in yellow) the ground glass opacities (GGO) over the original image (see Fig. 10) .

2.3 An index to define the severity of pneumonia

One main feature/index can be given to the radiologist as a measure of the severity of the infection: the percentage of ground glass opacities in the lungs.

The objective of this feature is to give the doctor an index representing more or less the amount of ground glass opacities in the lungs. With respect to Fig. 10, for the chosen slice or section of the body (slice 133), it was measured an overall percentage (with respect to the area of both lungs) of around 14% of ground glass opacities.

Focusing on each lung, the left one resulted to have a percentage of the infected area of almost 15%, while for the right lung this value was equal to almost 13%.

The higher the value of this feature, the worst should be the patient status as a higher amount of lung alveoli is partially filled with exudate or is partially collapsed and the tissue around them is thickened.

The indexes show how the left lung is in a worse condition than the right one.

However, this index is provided from the algorithm to the doctor, who should draw his own conclusions in terms of the severity of the illness.

3 Conclusions

Pneumonia is an inflammatory condition of the lung primarily affecting the small air sacs known as alveoli. Symptoms typically include some combination of productive or dry cough, chest pain, fever, and difficulty breathing [4]. The severity of the condition is variable. It is believed that the pulmonary lesions caused by COVID-19 infection are similar to those of pneumonia.

CT scans produce 2-dimensional images of a "slice" or section of the body, but the data can also be used to construct 3-dimensional images [5]. A CT scan of the chest may be done to see finer details within the lungs and detect pneumonia that may be more difficult to see on a plain x-ray. A CT scan also shows the airway (trachea and bronchi) in great detail and can help determine if pneumonia may be related to a problem within the airway. A CT scan can also show complications of pneumonia, abscesses or pleural effusions and enlarged lymph nodes [6].

As it can be seen in COVID-19 CT scans from [1], there is the presence of ground glass opacities-GGO ("frosted" areas) with a lighter grey color inside the lungs (see Fig. 1). The grey areas indicate increased density, meaning that something is partially filling the air spaces inside the lungs. This could be due to the air spaces becoming partially filled with fluid (also pus or cells), the walls of the alveoli (which are the tiny air sacs in the lungs) thickening, or the space between the lungs thickening.

GGO can be due to many conditions. Sometimes, the cause is benign but other times it may be the temporary result of a short-term illness. However, it can also indicate a more serious or long-term condition.

A 2020 study [7], in which CT scans of patients affected by COVID-19 were analyzed, found that GGO most commonly showed up in the lower lobes of the lungs as round opacities, but that as the disease progressed, it became more patchy and affected all lobes.

As it is shown in Fig. 10, for the chosen CT scan slice (slice 133), the left lung seems to be in a worse condition than the right one as there is a higher distribution of GGO.

It also can be considered that it should be better to have denser and localized distributions of ground glass opacities in specific zones of the lung rather than having a scattered distribution of ground glass opacities covering the whole lung area.

The aim of the algorithm implemented is to isolate the ground glass opacities inside the lungs of patients affected by COVID-19. The final result is shown in Fig. 10. The parameters used to generate the infection mask were: the range of values ([-700, -250]), the size of the square kernel (5x5) used in filtering the image, the threshold to detect opacity (0.2). Since the GGO are grey areas, in the infection mask generated by the algorithm they should be composed of several yellow points (see Fig. 10), hence the set of the parameters was made in order to reduce the number of isolated yellow points which do not identify GGO.

In conclusion, the final result should be as realistic as possible, hence reliable for the doctor, who should only analyze it.

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] `https://en.wikipedia.org/wiki/Pneumonia`

[5] `https://www.medicalnewstoday.com/articles/153201`

[6] `https://www.radiologyinfo.org/en/info/pneumonia`

[7] `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7360078/`