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

Marco Colocrese, s301227,
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.

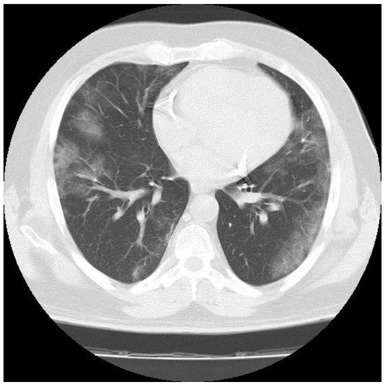

Figure 1: Example of ground glass opacity (light grey opaque areas in the lungs).

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.

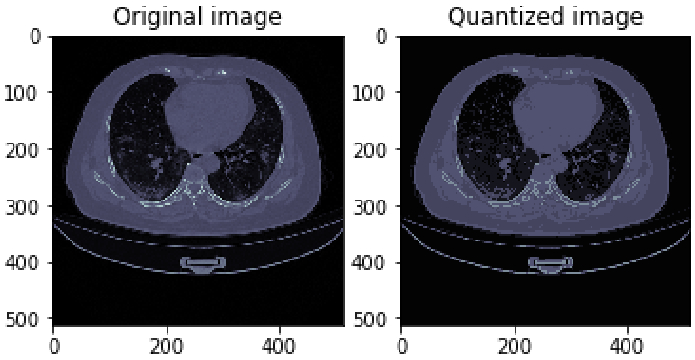

Figure 2: Original (left) and color quantized (right) images.

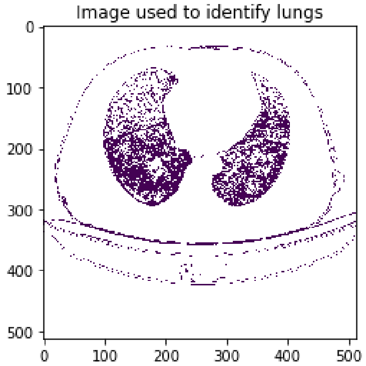

Figure 3: Region with the second darkest color after quantization through K-means.

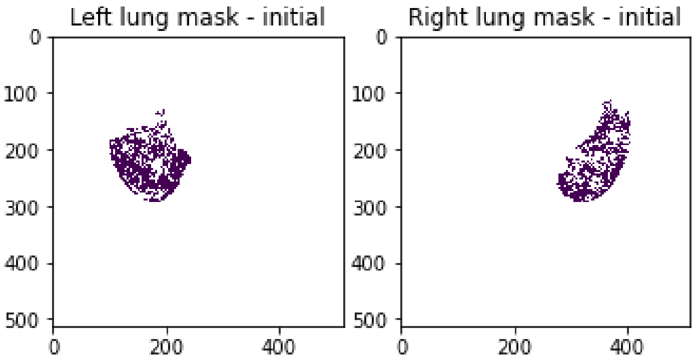

Figure 4: Initial identification of lungs.

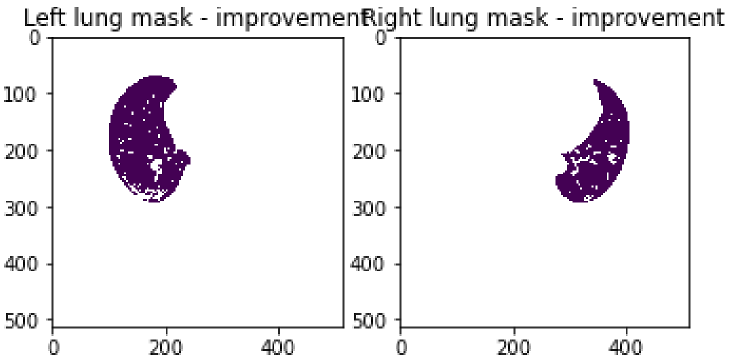

Figure 5: intermediate identification of lungs.

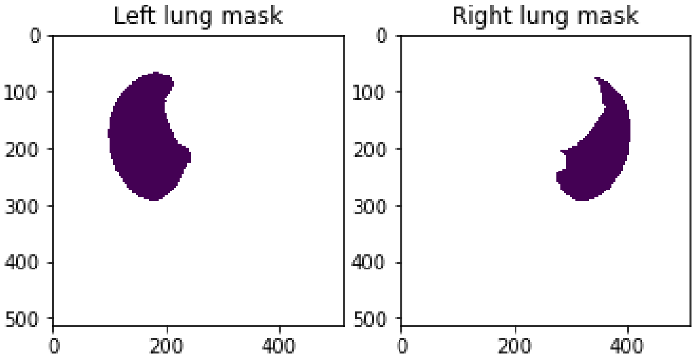

Figure 6: Final identification of lungs.

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 right; this imperfection is 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 $[-650, -300]$ to detect pixels corresponding to possible

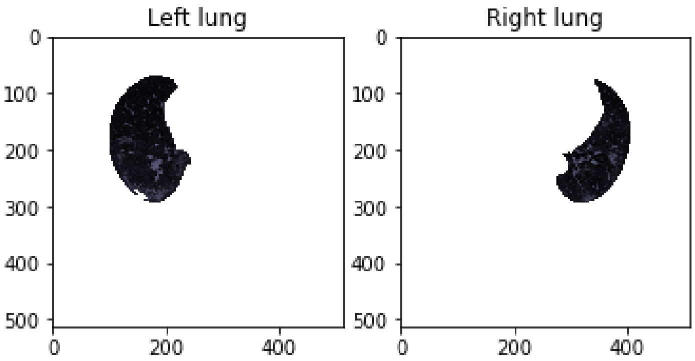

Figure 7: Image of lungs with bone colormap.

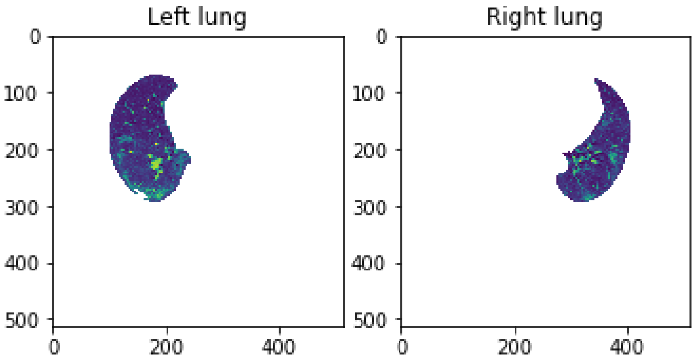

Figure 8: Image of lungs with viridis colormap.

infection area; then filtering was used with a kernel with size 3x3 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.

Fig. 9 show the detected inflamed points and Fig. 10 represents those overlapped to the original figure showing the goodness of the software that stressed the infected zones. The data plotted in these figures were then analyzed to produce the output features. The figure in the right of Fig. 10 is the same as on the left with some adjustments obtained by deleting some outliers. To do this, DBSCAN was used with parameters epsilon=5 and minimum samples=5. As it can be noticed, not all the outliers were deleted, as DBSCAN parameters were chosen conservatively, in order to avoid the deleteing of useful points. In any case, both the plots are reported so that the radiologist can evaluate by himself.

2.3 Some indexes to define the severity of pneumonia

To define the severity of pneumonia the number and the distribution over the lungs of the infected zones (with the unit of measure equal to 1 pixel) were analyzed. The software gives as output:

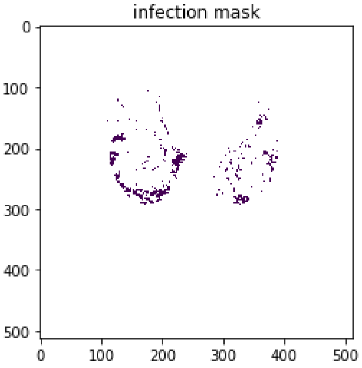

Figure 9: Infection mask.

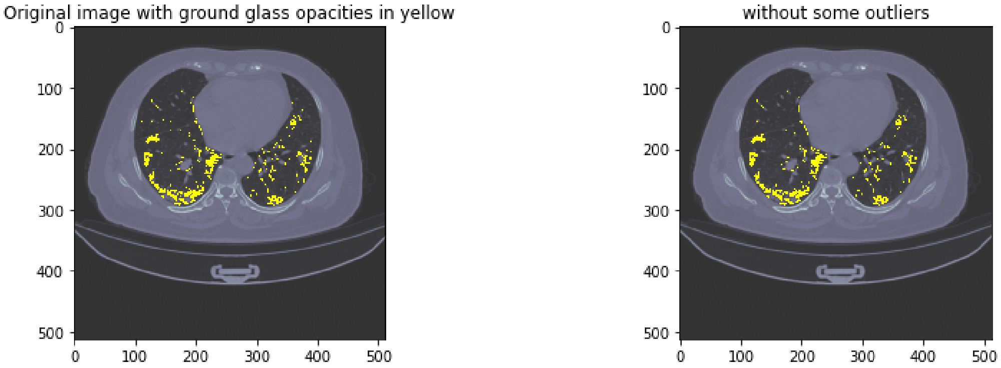

Figure 10: Infection region (in yellow) superimposed to the original image.

- the percentage of 'infected pixels' considering both the lungs;
- the percentage of 'infected pixels' of each lung (note that the previous output is not the mean of these two percentages as the lungs have different dimensions);
- the number of big, medium and small zones. To compute these statistics DBSCAN was used with parameters eps=5 and minimum samples=2. The dimensions of the obtained clusters were analyzed to count the number of big clusters (at least 3% of the pixels of the considered lung), medium clusters (number of pixels between 0.5% and 3% of the total of the lung) and small clusters (clusters with number pf pixels smaller than 0.5% of the total of the lung and outliers).

In the reported example these features are:

- overall infected lungs percentage: 8.5%;
- infected left lung percentage: 9.9%, infected right lung percentage: 6.6%;
- left lung infection distribution: 1 big zone, 3 medium zones and 38 small zones; right lung infection distribution: 0 big zones, 3 medium zones and 44 small zones.

This output is thought to be used by the medical doctor together with the output Fig. 10. Overall, the result seems accurate: comparing it to the plot, the percentage and the zones are correctly represented.

3 Conclusions

With the spread of SARS-Cov-2 Virus Infection, the medical sector and the linked ones had to face different challenges. One of the most difficult was the identification of a fast, accessible and reliable diagnostic modality used as an alternative to RT-PCR. The analyzed Chest computerised Tomography (CT) can be considered a possible instrument in diagnosing COVID-19. Moreover it is possibly more reliable, useful and quicker than RT-PCR[4] and a study in Wuhan concluded that chest CT imaging demonstrated a sensitivity of 97% (using RT-PCR as the reference). A computerized tomography (CT) scan combines a series of X-ray images taken from different angles around your body and uses computer processing to create cross-sectional images (slices) of the bones, blood vessels and soft tissues inside your body. In the results of the example studied in this paper, only one slide is considered. CT scan images provide more-detailed information than plain X-rays do[5].

As pneumonia is one of the most harmful symptoms (together with severe acute respiratory syndrome and renal failure), it is also useful to have many information about the lungs health of the patient, this is linked to the output features choice. Output features also help to relate the detected pneumonia to its nature, for instance bacterial pneumonia usually affects only a part of one lung being dense, unlike of pneumonia of viral cause that can be scattered, affecting both lungs, especially in the lower lobes[6].

In general, ground-glass opacities and consolidations, with a bilateral and peripheral distribution, are the most typical patterns found in COVID-19 pneumonia[7].The typical imaging manifestations of early COVID-19 are patchy, rounded, segmental or subsegmental ground-glass opacities with or without consolidation. Lesions are multiple and asymmetrically distributed and are more common in the peripheral areas[8]. Moreover, a huge quantity of small infected zones can be linked to a more diffuse pneumonia that generally represents the worst situation.

In conclusion, this software can help medical doctors to have an objective and quantitative interpretation of CT scan to characterize lungs inflammation related to COVID-19; it can obviously used for generic applications of CT scan, not only for COVID-19 detection.

Through the use of the output features and the output plot, the CT scan result can be more easily understood, but a direct diagnosis of COVID-19 must be given by the medical doctor considering also other parameters such as other specific symptoms as SARS-CoV-2 is only one of the possible causes of pneumonia.

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://link.springer.com/article/10.1007/s10140-020-01886-y#ref-CR10

[5] https://www.mayoclinic.org/tests-procedures/ct-scan/about/pac-20393675

[6] Umberto Veronesi, *Salute per tutti - dizionario medico*, 2008

[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8567439/

[8] https://pubs.rsna.org/doi/full/10.1148/ryct.2020200047