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

Francesco Conforte, s277683,
ICT for Health attended in A.Y. 2020/21

December 3rd, 2020

## 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 the typical findings of suspected or confirmed cases of COVID-19 infection.

The common typical chest CT scan findings, can be detected through scans are summarized as: Peripheral distribution, Bilateral lung involvement, Multifocal involvement, Ground glass opacification (instead of appearing uniformly dark, show greyish areas with ground glass opacity (GGO)), 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.

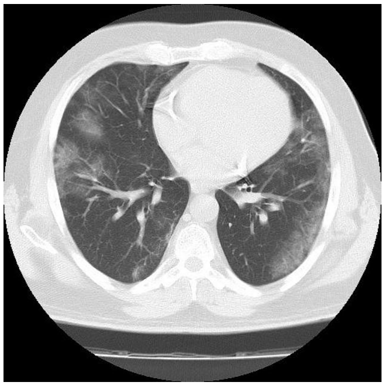

Figure 1: Example of ground glass opacity (light grey opaque areas in the lungs).

# 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 $512 \times 512$ 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 $512 \times 512$ 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 right and left 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

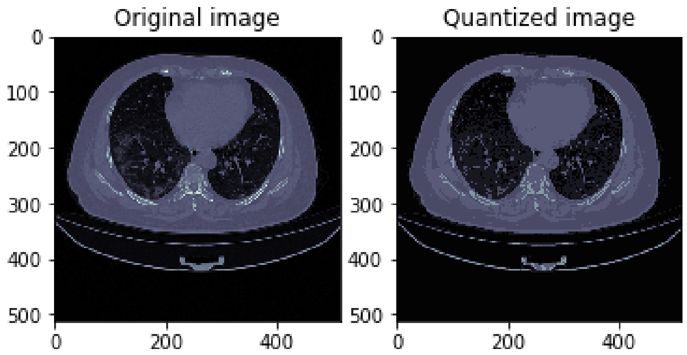

Figure 2: Original (left) and color quantized (right) images.

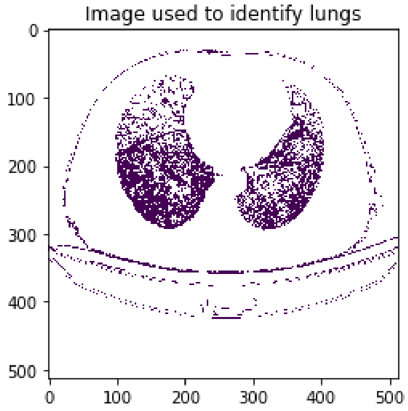

Figure 3: Region with the second darkest color after quantization through K-means.

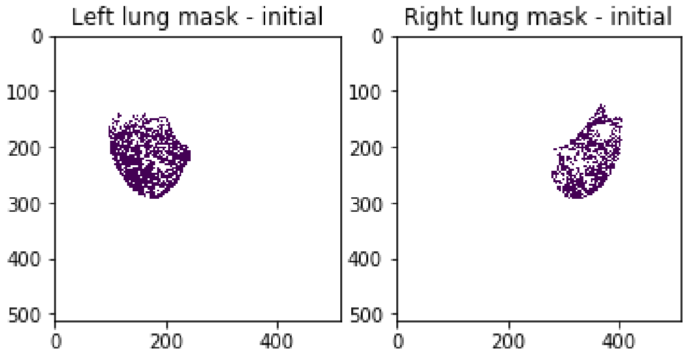

Figure 4: Initial identification of lungs.

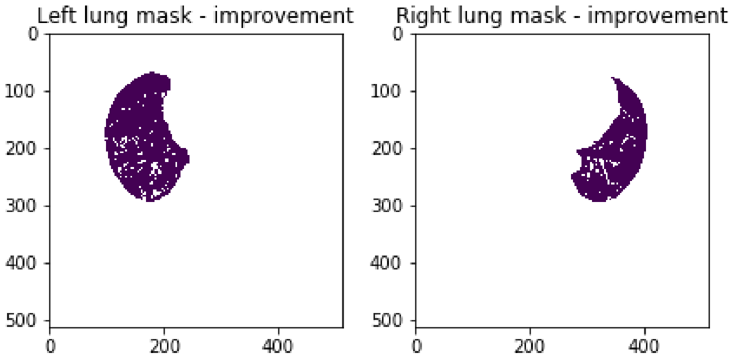

Figure 5: Intermediate identification of lungs.

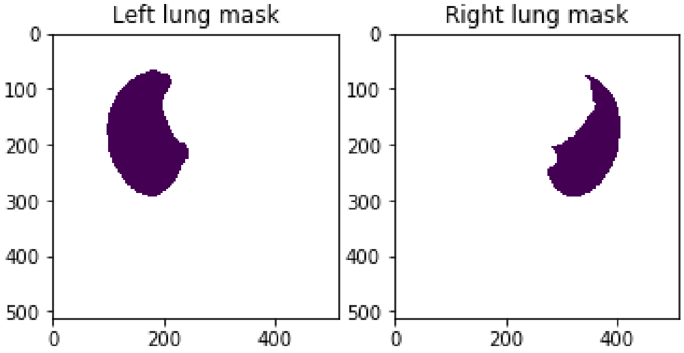

Figure 6: Final identification of lungs.

Fig. 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.

## 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 $[-650, -300]$ to generate the images in Figs. 9 and 10.

Fig. 9 shows the pixels corresponding to the infection mask for both lungs and Fig. 10 shows the original image which the infection mask is superimposed to. Needless to say, the applied process to find GGOs explained in this section is not an exact science, meaning that some opacities may also be missing in the recognition process but most of them are perfectly identified in both of lungs.

This work can be exploited to define a useful index to be used by radiologists as explained in Section 2.3.

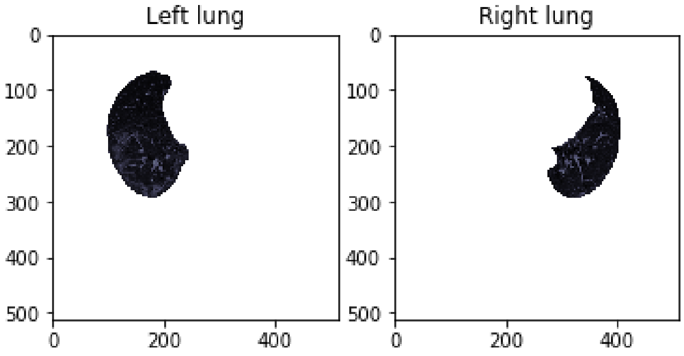

Figure 7: Image of lungs with bone colormap.

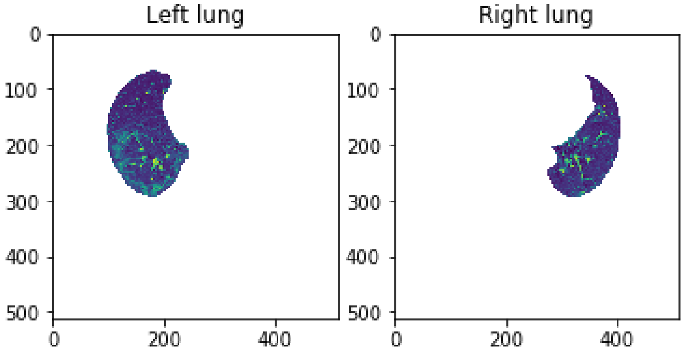

Figure 8: Image of lungs with viridis colormap.

## 2.3 An index to define the severity of pneumonia

In order to quantitatively estimate the pulmonary involvement, a **score** on a scale from 0 to 5 can be used on the basis of the involved lung area, with:

- 0 indicating **no** involvement;
- 1 indicating **less than 5%** involvement;
- 2 indicating a range of **5%–25%** involvement;
- 3 indicating a range of **26%–49%** involvement;

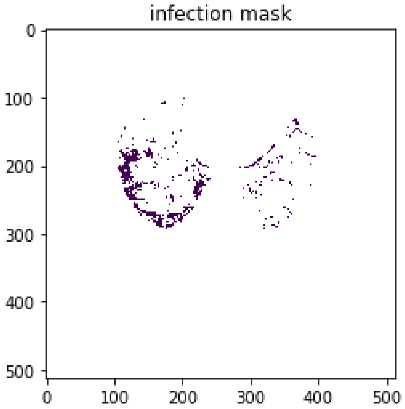

Figure 9: Infection mask.

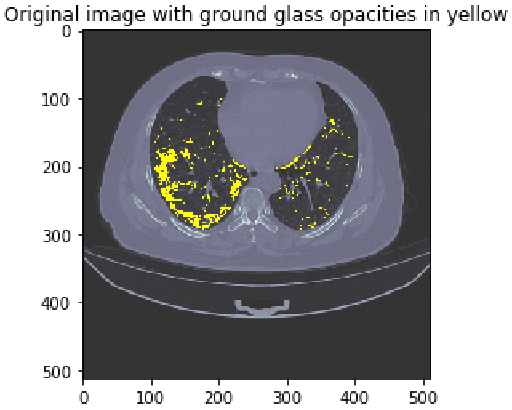

Figure 10: Infection region (in yellow) superimposed to the original image.

- 4 indicating a range of **50%−75%** involvement;
- 5 indicating **more than 75%** involvement.

Each score can be assigned to each of the five lung lobes (Right Superior Lobe, Right Middle Lobe, Right Inferior Lobe, Left Superior Lobe and Left Inferior Lobe) and the sum of the five assigned scores can be computed to get a single score, ranging from 0 (no involvement) to 25 (maximum involvement): the higher the score, the larger the distribution of lung abnormalities and, in turn, the severity of pneumonia.

For study purpose, the thought index was tested on a single slice of CT scan shown in Fig. 10, representing a slice of Right Middle Lobe and Left Inferior Lobe. Accordingly, the right lung is more infected with respect to the left lung, with a percentage of 13.26% versus a percentage of 4.58%. Therefore, the score assigned to the Right Middle Lobe was **2** and the one assigned to the Left Inferior Lobe was **1**.

In order to have a more efficient result from a medical point of view, the algorithm may be applied to each slice belonging to a set of slices corresponding to a lobe and the score may be assigned on the basis of an average computed over all the scores assigned to each slice.

Moreover, radiologists are more interested in observing results about the whole lung and not only a single slice of a single lobe. For this reason, it is suggested to compute firstly each score for each lobe and finally to sum the five scores.

# 3 Conclusions

Computing the index described in Section 2.3 may be very useful to determine not only the severity of pneumonia but also to follow how chest CT findings, associated with COVID-19, change from initial diagnosis until patient recover. This would be essential in order to establish stages of interstitial pneumonia: early stage, progressive stage, peak stage and absorption stage (see [4]).

Whilst chest CT scan may seem a rapid way to detect common features of pneumonia inferred by coronavirus disease-19, it is important to realize that CT is not the standard for the diagnosis of COVID-19. It is determinant to correlate chest CT findings with epidemiological history and, more importantly, with RT-PCR (Real-Time reverse transcription-Polymerase Chain Reaction) test results. Indeed, as highlighted in [5], the reported chest CT abnormalities in COVID-19 are similar to those seen in infections with SARS-CoV-1 and MERS-CoV as well as the ones caused by influenza type A. Therefore, the chest CT scan is only a further tool that doctors have at disposal to win this war with the common enemy known to all as COVID-19.

Finally, it is also important to claim that the usage of CT scan on patients cannot be abused, because of the radiation burden that a patient absorbs each time he undergoes a chest screening.

## References

[1] COVID-19 CT scans. 20 CT scans and expert segmentations of patients with COVID-19. `https://www.kaggle.com/andrewmvd/covid19-ct-scans`.

[2] K.P. Murphy. *Machine Learning: A Probabilistic Perspective.* Adaptive Computation and Machine Learning series. MIT Press, 2012.

[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and 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).* pages 226–231. AAAI Press, 1996.

[4] Feng Pan, Tianhe Ye, Peng Sun, Shan Gui, Bo Liang, Lingli Li, Dandan Zheng, Jiazheng Wang, Richard L. Hesketh, Lian Yang, and Chuansheng Zheng. Time Course of Lung Changes at Chest CT during Recovery from Coronavirus Disease 2019 (COVID-19). *Radiology*, 295(3):715–721, 2020. PMID: 32053470.

[5] Thomas C. Kwee and Robert M. Kwee. Chest CT in COVID-19: What the Radiologist Needs to Know. *RadioGraphics*, 40(7):1848–1865, 2020. PMID: 33095680.