COVID-19

Early detection in Chest X-ray images

Report of "Data distribution variability among acquisition devices"

April 9, 2020

Contents

1 Motivation

Chest x-ray radiograph (CXR) is becoming crucial for following the clinical evolution of covid19 patients. The ease of transport of the acquisition devices makes it more extended than the tomography, despite the image quality and the diagnostic performance are superior using such technology. The high number of detected covid19 cases around the world increases the necessity of an automatic tool to detect the pulmonary signs (pneumonia or infiltration) in the CXR from these patients, which in turn would reduce the specialist workload and allow them to focus their efforts in the most critical patients.

For tasks in which medical image is involved, the paradigm of deep learning (DL) has been demonstrated to be a powerful instrument [1]. But it has also been proved that the performance of the models based on this technology may be improved by the use of some image processing steps [2, 3, 4, 5].

With the purpose of pre-training models and getting experience in the aim to detect covid19 cases, we will use DL models to classify patients in **Controls (C)**, **Pneumonia (P)**, **Infiltration (I)** or **Pneumonia and Infiltration (PI)**. These models will be updated to fit the original task (covid19 detection), when covid19 CXRs are available.

Since the current challenge is relevant around the world, the acquisition devices used for the follow up of the patients will probably present significant differences in different countries due to the different quality of the acquisition devices which can exist in the clinical practice from different continents. In this sense, to investigate methodologies to decrease that acquisition variability becomes critical. The current document presents a histogram analysis of the images to study the differences among data sources.

2 Methods

A total of 23520 CXR coming from Medical Imaging Databank of the Valencia Region (BIMCV) repository were used to evaluate the histogram distribution among image modality (Computerized Radiography (CR) and Digital Radiography (DX)), along with image acquisition device manufacturer (Philips Medical Systems and Imaging Dynamics Company Ltd). Besides, thanks to the work conducted by Cohen et al. [6], it was possible to extract 162 images from covid19 patients to also assess and compare their histogram distribution.

In CXR images that have been converted to 8-bit pixel representation, each pixel x of the image verifies that $x \in [0, 255] \subset \mathbb{Z}$. By building the 256-bins histogram from each image, and using the widespread *Jensen-Shannon* distance (JSD) (see Equation 1), we are capable to define a simplex, where each point represents the histogram of an image, the distance between two points is the JSD between the two histograms represented by the points. Furthermore, this simplex is contained in a *statistical manifold*, which is in turn a Riemannian manifold, and its dimension is at most $n-1$ where n is the number of images.

$$JSD(P,Q) = \sqrt{\frac{\sum_i P(i) \log \frac{P(i)}{M(i)} + \sum_i Q(i) \log \frac{Q(i)}{M(i)}}{2}} \text{ where } M = \frac{P+Q}{2} \tag{1}$$

After applying dimensionality reduction methods such as Principal Component Analysis (PCA), it is possible to obtain visual representations of the histogram's distribution which not only allows us

to find faulty images but also to assess differences among data sources.

3 Results

After applying the aforementioned methodology to the BIMCV CXR database, the distribution of the histograms presents a bimodality as can be seen in Figure 1. Furthermore, the images represented by the points inside the red rectangle resulted to be completely faulty, meanwhile many of the images represented by the points inside the blue rectangle resulted to present a very large black or white frame around the chest.

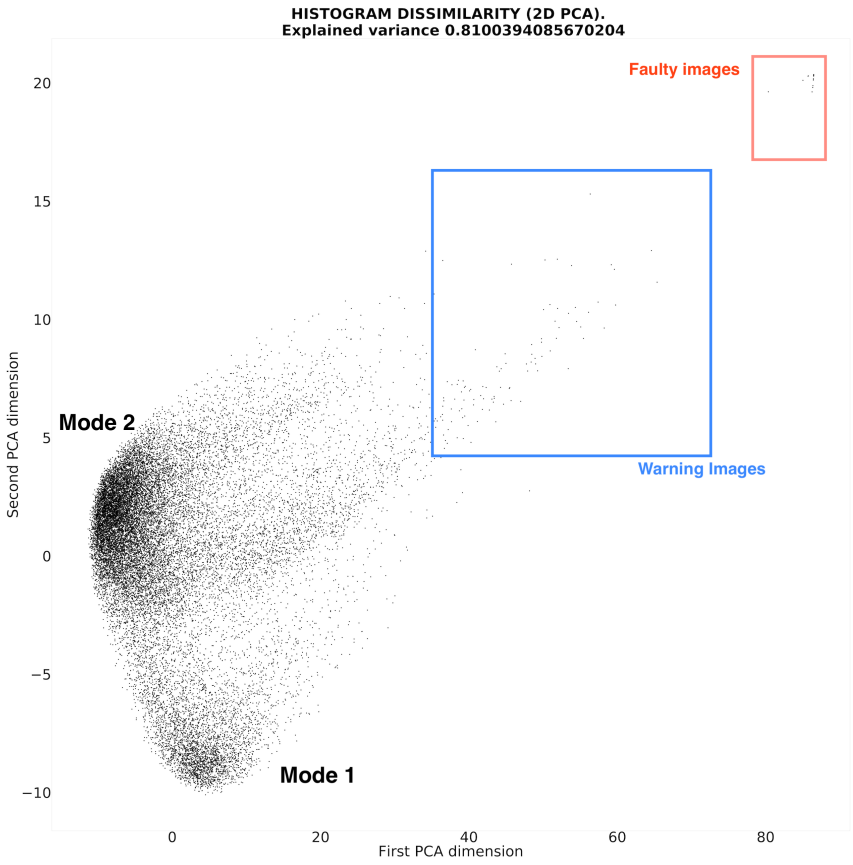

Figure 1: Histogram dissimilarity. The explained variance ratio of 0.81 over 1 indicates a good dissimilarity representation.

The BIMCV database was collected using two different acquisition modalities, CR and DX. The

Figure 2 shows the distribution of the histograms where the points differentiated by color according to its modality. The bimodality is not entirely explained by the modality of the images.

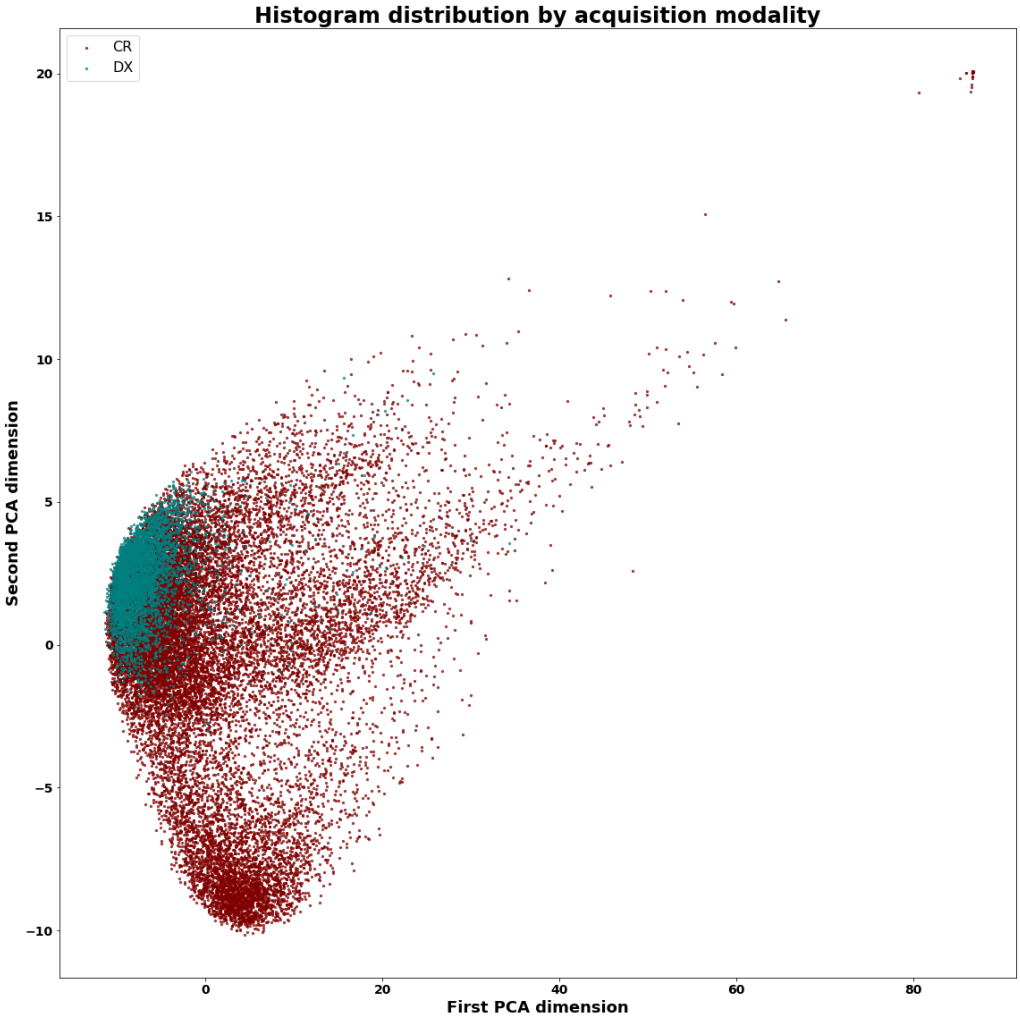

Figure 2: Histogram dissimilarity per acquisition modality. Blue points correspond to DX images and red points correspond to CR images.

The image metadata also provided the manufacturer company of the acquisition devices, Figure 3 shows the distribution of the histograms were the points differentiated by color according to the manufacturer of the acquisition device. The bimodality of the images analyzed is most probably due to the way each manufacturer processes the image.

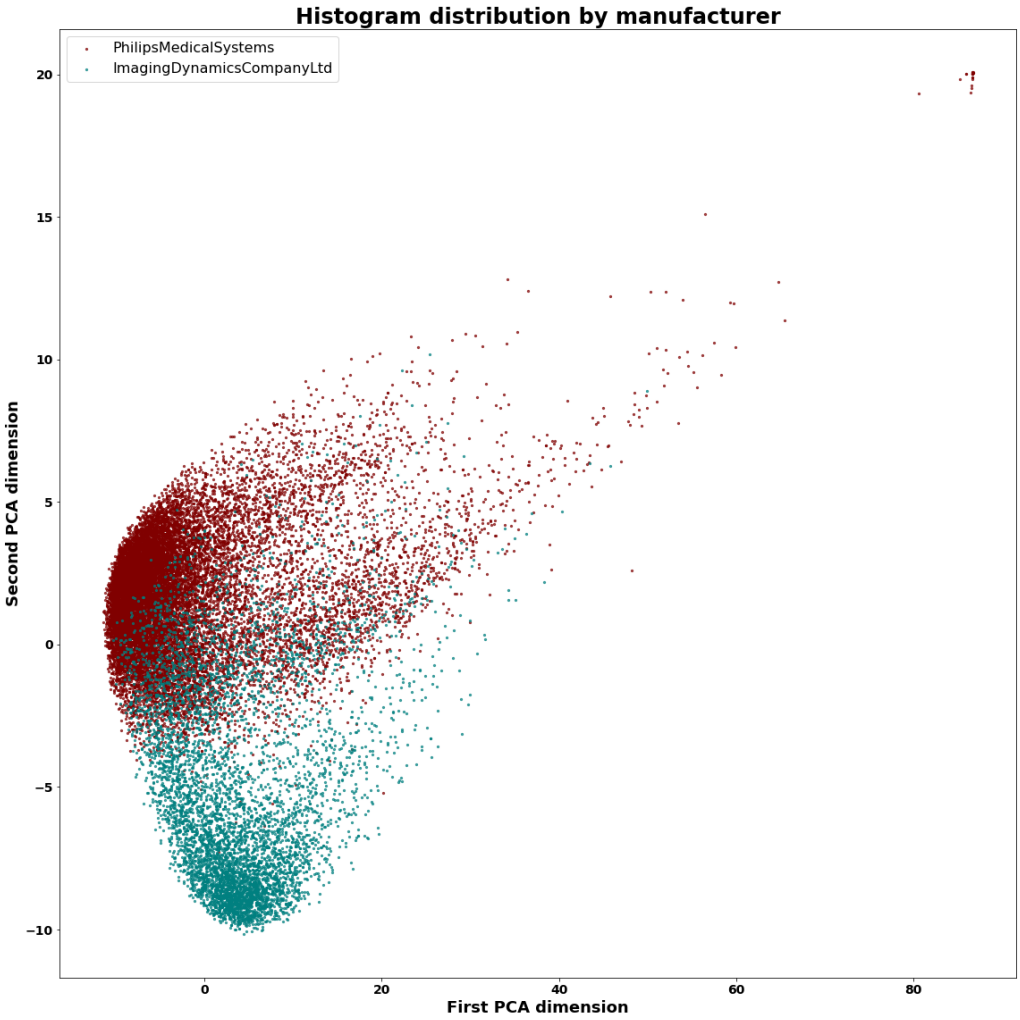

Figure 3: Histogram dissimilarity per acquisition device manufacturer. Blue points correspond to PhilipsMedical Systems and red points correspond to Imaging Dynamics Company.

We have also verified that the histograms of the images of the training, validation and test sets are equally distributed across the simplex, as can be seen in the Figure 4.

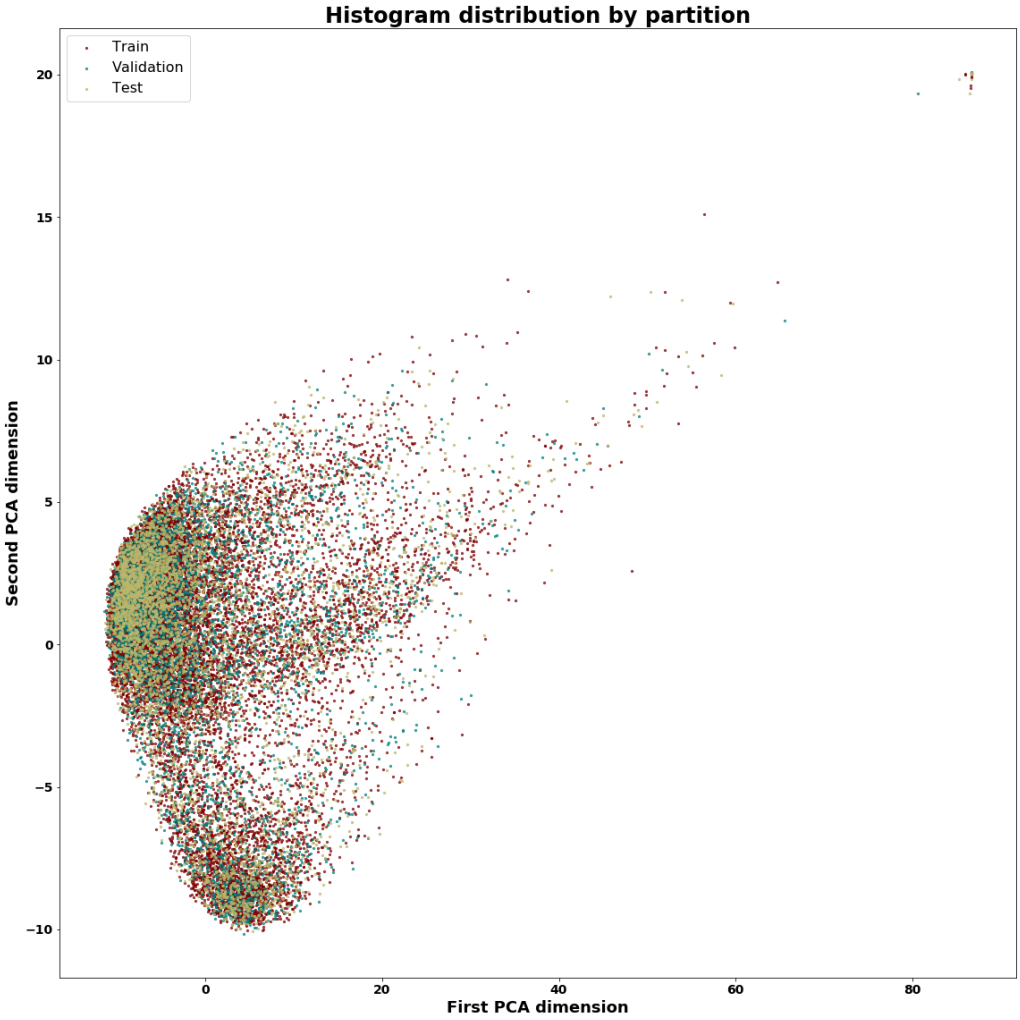

Figure 4: Histogram dissimilarity per model partition. Red points correspond to the images chosen for the training set, blue points correspond to images that belong to the validation set and green points are those images selected to test the performance of the models.

Finally, since some covid19 images were available, the analysis was repeated after adding these images. Figure 5 shows that the histograms of these images are compatible with the distribution of the previous ones.

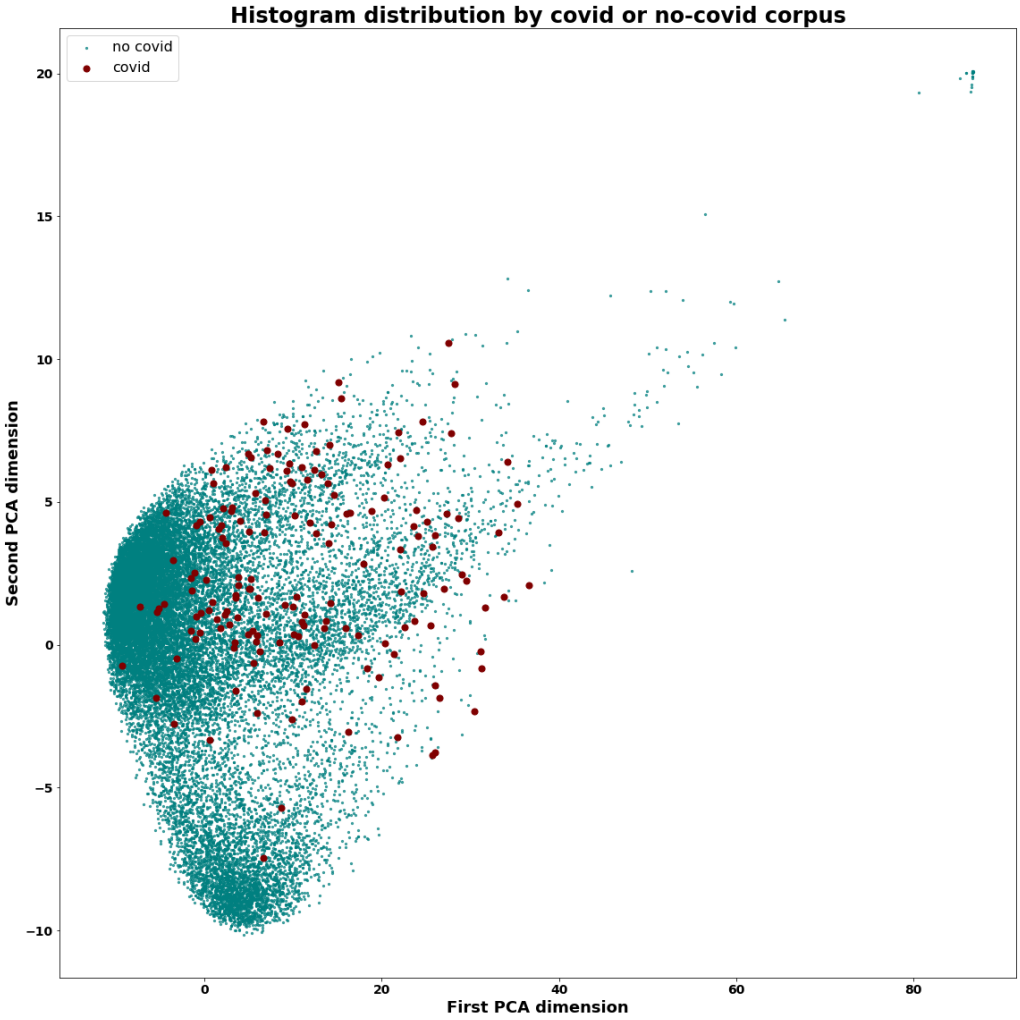

Figure 5: Images from ieee8023 github corresponding to **COVID-19** patients were collected. Red points are the histograms for covid19 patients. It shows the dissimilarity of their histograms are comparable to the images from the original dataset.

4 Conclusions

This report shows the results of an analysis based on grey-value histogram information showing the differences among data acquisition sources in medical imaging, in particular CXR images. This work also motivates the use of imaging preprocessing to decrease the variability of images, which may lead to a better performance of DL models. We encourage researchers to define methods to reduce the differences among data sources. To know the acquisition device may help to define images registration methods for each acquisition device, but this information will not always be available; in this sense, we propose to define the registration from a prototype or "average histogram" of the observed images.

Acknowledgments

Data provided by the Banco de Imágenes Médicas de la Comunitat Valenciana and several international image banks, with the aim of having the widest possible sample. This report has been carried out by the Instituto Tecnológico de Informática (ITI), in a joint collaboration of the Universidad Politécnica de Valencia (UPV) and the participation in the dataset provision, preparation and colaborative analysis of the Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO), the Universidad Miguel Hernández, the Universidad de Alicante, staff from the Hospital San Juan de Alicante, the Centro de Investigación Príncipe Felipe and the companies MedBravo and GE. The initiative is promoted by the Conselleria de Innovación, Universities, Ciencia y Sociedad Digital, and the Conselleria de Sanidad Universal y Salud Pública.

References

[1] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, 2016. 1

[2] R. M. Sundhari, "Enhanced histogram equalization based nodule enhancement and neural network based detection for chest x-ray radiographs," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–9, 2020. 1

[3] Y. Gordienko, P. Gang, J. Hui, W. Zeng, Y. Kochura, O. Alienin, O. Rokovyi, and S. Stirenko, "Deep learning with lung segmentation and bone shadow exclusion techniques for chest x-ray analysis of lung cancer," in *International Conference on Computer Science, Engineering and Education Applications*, pp. 638–647, Springer, 2018. 1

[4] A. K. Bhandari, "A logarithmic law based histogram modification scheme for naturalness image contrast enhancement," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 4, pp. 1605–1627, 2020. 1

[5] R. Sivaramakrishnan, S. Antani, S. Candemir, Z. Xue, J. Abuya, M. Kohli, P. Alderson, and G. Thoma, "Comparing deep learning models for population screening using chest radiography," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575, p. 105751E, International Society for Optics and Photonics, 2018. 1

[6] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv 2003.11597*, 2020. 2