CT normalization by paired image-to-image translation for lung emphysema quantification

Insa Lange[1], Fabian Jacob[2], Alex Frydrychowicz[2], Heinz Handels[1], Jan Ehrhardt[1]

[1]Institute of Medical Informatics, University of Lübeck
[2]Department for Radiology and Nuclear Medicine, University Hospital of Schleswig-Holstein
insa.lange@student.uni-luebeck.de

Abstract. In this work a UNet-based normalization method by paired image-to-image translation of Chest CT images was developed. Due to different noise-levels, emphysema quantification shows sincere subordination to the choice of the filterkernel. Images for training and testing of 71 patients were available, reconstructed using the smooth Siemens B20f filterkernel and the sharp B80f filterkernel. Results were evaluated in regard to the image quality, including a visual assessment by two imaging experts, the L1 distance, the emphysema quantification (emphysema index and Dice overlap of emphysema segmentations). Emphysema quantification was compared to classical normalization methods. Our approach lead to very good image quality in which the mean B20f L1 distance to the B80f could be reduced by about 88.5 % and the mean Dice was raised by 189 % after normalization. Classical methods were outperformed. Even though small differences between B20f and normalized B80f images were noticed, the normalized images were found to be overall of diagnostic quality.

1 Introduction

Chronic-obstructive pulmonary disease (COPD) is a very common widespread illness that is responsible for many deaths worldwide. Patients suffering from COPD are experiencing a limitation of airflow to the lungs often caused by an alveolar obstruction called lung emphysema. Thus, diagnosing COPD and evaluating progress of the disease is of great significance. Standard diagnostic methods are pulmonary function tests using spirometry, however, computed tomography (CT) is increasingly used to exactly locate affected areas in the lung.

Quantification of lung emphysema in Chest-CTs can be achieved by a simple threshold segmentation of emphysema areas inside the lung, meaning lung segmentations are needed. A widely accepted threshold is $< -950\,HU$ and a global emphysema index (EI) can be computed from the segmented emphysema volume (EV) and the total lung volume (TLV) by $EI = \frac{100\% \times EV}{TLV}$. Inconveniently, the calculated emphysema index strongly depends on the selected CT

imaging parameters such as slice thickness or reconstruction kernel, which makes CT normalization methods necessary to achieve comparable results [1,2].

This paper focuses on CT normalization with respect to the chosen reconstruction kernel. EI quantifications based on images reconstructed using smooth kernels (like Siemens B20f) seem to correspond well to pulmonary function tests [2]. In comparison, quantifications based on images reconstructed by sharp kernels (like Siemens B80f) will result in deviation of the EI, suggesting a much greater EV is present than it actually is. Previous work suggested filtering methods [3,4] or frequency decomposition methods [2,4], which didn't take the truncation artifact at $-1024\,HU$ into account, to normalize CT images reconstructed with different kernels to a consistent appearance. A more recent approach successfully applies image-to-image translation (Siemens: B50f $\rightarrow$ B30f) by Convolutional Neural Networks (CNNs) for this task [5].

The aim of this work is to investigate the possibilities and limitations of state-of-the-art image-to-image translation techniques for CT kernel normalization. For this purpose, we consider the challenging task of normalizing Siemens B80f images (very sharp kernel) and translating the images into B20f images (very smooth kernel). Our work should contribute to the question whether images generated in this way are suitable for clinical diagnostics. Therefore, we present a quantitative evaluation of the generated images and the calculated EI and Dice, as well as a visual assessment of image quality by an experienced radiologist and an in medical imaging specialized computer scientist.

2 Materials and methods

2.1 Materials

Imaging data of 71 patients were acquired on a 64-row MSCT (Somatom Definition AS+, Siemens) and reconstructed simultaneously with a soft (B20f) and a sharp kernel (B80f). Slice thickness varied between $0.7\,mm$ and $1\,mm$ and was resampled to $1\,mm$ for all reconstructions. The accustomed in-slice spacing and slice dimensions of 512×512 pixels were kept unchanged. Due to the simultaneous reconstruction, the images form a well suited basis for paired image-to-image translation techniques.

2.2 Image-to-image translation with UNet

In the past, both paired image-to-image translation, e.g. Pix2Pix [6], MedGAN [7] or MEGAN [8], and unpaired image-to-image translations, e.g. CycleGAN [9] produced remarkable results. Paired image translation methods apply pixel-wise error losses to corresponding images of each domain, whereas unpaired methods such as CycleGAN typically use distribution matching losses that often result in hallucinated features in the generated images [10]. This effect has been less observed in paired methods [10], and we believe that paired image-to-image translation with the same raw data for each image pair can be used for diagnostic purposes.

For the paired image-to-image translation of B80f images to B20f images a UNet architecture with three skip-connections, maxpooling for compression and 128 feature maps in maximal compression was build and multiple times trained in 100 epochs. In each training eleven to twelve patients were left out for testing and eleven to twelve patients for validation. Due to the great size of the CT data, images were cut into patches for training. For all of the chosen patients to be used in training the same number of randomly selected non-overlapping patches was added to the training dataset. L1 loss was used to compare to the original B20f.

2.3 Experiments

The quantitative analysis compares the original B20f images to the normalized B80f images. The differences in HU values (L1 distance), and computed EI, as well as the Dice overlap between the emphysema regions are computed. Our UNet-based approach is compared to a heuristic filter-based approach [4] and a frequency decomposition method [2]. Paired one-sided T-tests were performed.

For visual evaluation, either an original B20f image or a normalized B80f image was randomly selected. Two experts rated the selected images blinded: a clincal radiologist (>5 years experience) and a computer scientist specialized on medical image computing (>5 years experience). The radiologist used a radiological workstation with quality assured monitor for the evaluation. All images were assigned a score by both experts seperately, ranging from 0 (certain synthetic) to 5 (certain original), thus including three levels of certainty for each originals and synthetic images and leaving no rating score for undecided.

3 Results

Multiple CNN trainings were performed, so that each patient was part of the test set at some point. Fig. 1 shows an original B20f, original B80f and the normalized

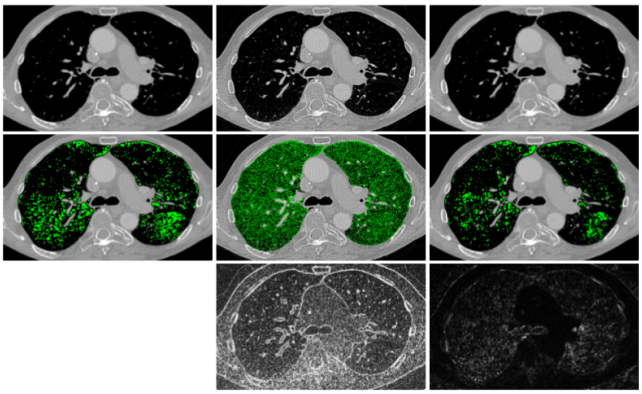

Fig. 1. (left to right) original B20f, original B80f, normalized B80f; (top to bottom) image, image with emphysema segmentation, difference image to original B20f.

Table 1. Statistics of Dice and absolute differences in EI.

Images	abs. difference in EI mean ± std, [min, max]	Dice mean ± std, [min, max]
B80f	17.32 ± 4.72, [5.78, 25.66]	0.19 ± 0.19, [0.01, 0.69]
UNet	0.93 ± 1.27, [0.01, 5.53]	0.55 ± 0.21, [0.03, 0.89]
Heuristic	1.90 ± 2.65, [0.04, 9.26]	0.43 ± 0.21, [0.00, 0.85]
Frequency-based	1.66 ± 2.26, [0.02, 9.36]	0.38 ± 0.25, [0.00, 0.85]
B80f, EI>5	17.17 ± 5.49, [5.78, 25.66]	0.41 ± 0.12, [0.25, 0.69]
UNet, EI>5	2.02 ± 1.45, [0.25, 5.53]	0.76 ± 0.07, [0.64, 0.89]
Heuristic, EI>5	4.31 ± 2.92, [0.39, 9.26]	0.63 ± 0.11, [0.47, 0.85]
Frequency-based, EI>5	3.44 ± 2.72, [0.35, 9.36]	0.61 ± 0.15, [0.19, 0.85]

B80f image, as well as emphysema segmentations and resulting difference images of one patient. The normalized B80f image is barely distinguishable from the original B20f and the typical over-segmentation of emphysema regions in B80f images is greatly reduced by the normalization. The computed EI for this patient is 17.45% (B20f), 34.76% (B80f) and 16.90% (normalized B80f), and computed Dice scores are 0.43 (B20f vs. B80f) and 0.73 (B20f vs. normalized B80f) showing an improved emphysema detection in the normalized images.

3.1 Quantitative results

The L1 distances averaged over all 71 patients and all voxels showed a remarkable reduction from (mean ± std) 84.5 HU ± 11.7 for B80f images to 9.7 HU ± 1.7 for normalized B80f images in comparison to the B20f images. Due to difficulties in the lung segmentation task, emphysema quantification was accomplished for only a subset of 50 patients. Emphysema quantification after UNet normalization was compared to the results of frequency decomposition and heuristic normalization.

While plotting the differences in EI calculations in Fig. 2 with and without normalization is barely showing any distinguishment between all three normalization methods, statistical calculations in Tab. 1 show an advantage of the UNet normalization method ($p<0.05$ compared to frequency-based normalization and $p<0.001$ for heuristic normalization and B80f).

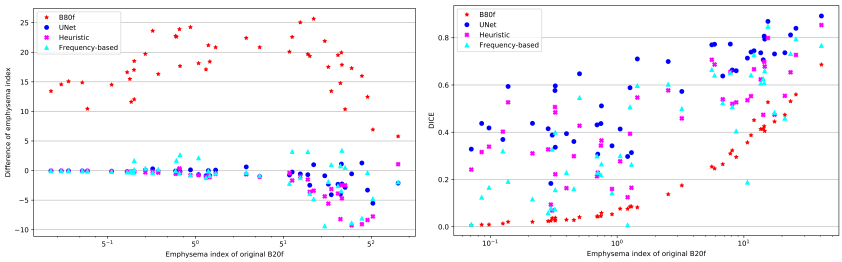

Fig. 2. (left) Differences in EI compared to groundtruth, (right) Dice values.

Looking at the Dice plot in this figure as well as the calculations in the same table, the UNet takes an obvious lead ($p<10^{-12}$ compared to both methods and B80f). Values suggest, that classical methods find a good EI compared to the baseline EI from B20f images. But as for exactly locating the airflow obstructions, those methods are not helpful and don't seem to gain any more information compared to performing a simple spirometry.

3.2 Visual assessment by experts

For 36 patients randomly either the original B20f or the normalized B80f image was selected, resulting in 16 original and 20 normalized images to be reviewed by the experts. Overall no image received a score of 0 or 5 to be certain original or synthetic. Mean score values and standard deviation for each group and both imaging experts can be reviewed as well as an overall score distribution can be reviewed in Fig. 3.

4 Discussion

This work presents a CNN-based paired image-to-image translation to normalize CT images reconstructed with different kernels (B80f→B20f). The aim of this work is to provide a CT normalization method for lung emphysema quantification and to evaluate its diagnostic usability.

The quantitative evaluation of the L1 distances in Hounsfield units and of the differences in the computed EI show a remarkable agreement between the baseline B20f images and the normalized B80f images. Thus, average differences in the computed EI are reduced from 17.61 before normalization to 0.98 afterwards. However, the average Dice score shows only a moderate agreement of 0.58 between baseline and normalized images. It must be noted, though, that for patients with a low EI (≤ 5), the segmented emphysema regions consist of only a few isolated pixels. For patients with medium to high EI, moderate to very good agreements are observed (Dice: 0.64–0.89). The normalization approach by UNet clearly outperformed the heuristic and frequency-based normalization methods [2,4] in regard to emphysema quantification, especially in Dice values.

The visual qualitative assessment of randomly selected original B20f and normalized B80f images has shown that neither the radiologist nor the computer scientist was able to distinguish between both image groups with absolute

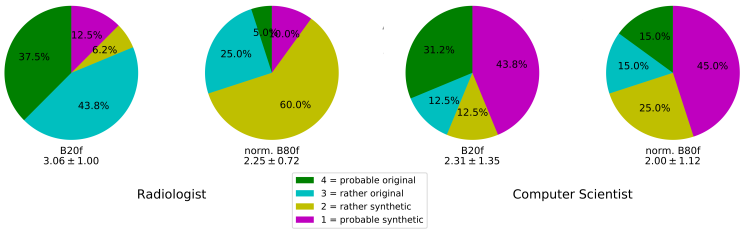

Fig. 3. Score distribution and mean ± std scores of visual assessment.

certainty. Although the computer scientist has many years of experience with medical image data, he could hardly distinguish between original and synthetic image data (58% of images correctly classified), whereas the radiologist tended to be able to differentiate between synthetic and original data (75% of images correctly classified). According to the radiologist, slight differences in the images can be detected at the emphysema borders.

In our opinion, a key lesson from this experiment is that a qualitative assessment of synthetically generated medical images should always be performed by experienced radiologists at clinical workstations. Nevertheless, according to an initial assessment by the radiologist, the normalized B80f images meet the diagnostic quality standards in radiology. A more in-depth analysis is necessary.

Even though difficult conditions for this research were chosen in translating very sharp and noisy images to very smooth images, high-quality images could be generated with a simple U-Net architecture with only three skip connections and a moderate size of the training set. A limitation of this study is that all image data for training and testing were acquired with the same type of CT scanner, so that for other settings new training data has to be acquired and the net has to be re-trained.

Building upon this research further CNN training with adjusted architecture can be expected in the attempt to minimize the differences in B20f and synthetic B20f images even more.

References

1. Bartel ST, Bierhals AJ, Pilgram TK, et al. Equating quantitative emphysema measurements on different CT reconstructions. Med Phys. 2011;38(8):4894–4902.
2. Gallardo-Estrella L, Lynch DA, Prokop M, et al. Normalizing computed tomography data reconstructed with different filter kernels: effect on emphysema quantification. Eur Radiol. 2016;26(2):478–486.
3. Ohkubo M, Wada S, Kayugawa A, et al. Image filtering as an alternative to the application of a different reconstruction kernel in CT imaging: feasibility study in lung cancer screening. Med Phys. 2011;38(7):3915–3923.
4. Ehrhardt J, Jacob F, Handels H, et al. Comparison of Post-Hoc Normalization Approaches for CT-based Lung Emphysema Index Quantification. Proc BVM. 2016; p. 44–49.
5. Jin H, Heo C, Kim JH. Deep learning-enabled accurate normalization of reconstruction kernel effects on emphysema quantification in low-dose CT. Phys Med Biol. 2019;64(13).
6. Isola P, Zhu JY, Zhou T, et al. Image-To-Image Translation With Conditional Adversarial Networks. Proc CVPR. 2017; p. 1125–1134.
7. Armanious K, Jiang C, Fischer M, et al. MedGAN: Medical image translation using GANs. Comput Med Imaging Graph. 2020;79.
8. Uzunova H, Ehrhardt J, Handels H. Memory-efficient GAN-based domain translation of high resolution 3D medical images. Comput Med Imaging Graph. 2020;89.
9. Zhu JY, Park T, Isola P, et al. Image-to-Image Translation with Conditional Adversarial Networks. Proc ICCV. 2017; p. 2223–2232.
10. Cohen JP, Luck M, Honari S. Distribution Matching Losses Can Hallucinate Features in Medical Image Translation. Proc MICCAI. 2018;11070:529–536.