

Comparison between threshold-based and deep learning-based bone segmentation on whole-body CT images.

Noémie Moreau^{a,b}, Caroline Rousseau^{c,d}, Constance Fourcade^{a,b}, Gianmarco Santini^b, Ludovic Ferrer^{c,d}, Marie Lacombe^d, Camille Guillerminet^d, Pascal Jezequel^c, Mario Campone^{c,d}, Nicolas Normand^a, and Mathieu Rubeaux^b

^aUniversité de Nantes, CNRS, LS2N, F-44000 Nantes, France

^bKeosys Medical Imaging, Nantes, France

^cUniversity of Nantes, CRCINA, INSERM UMR1232, CNRS-ERL6001, Nantes, France

^dICO Cancer Center, Nantes - Angers, France

1. DESCRIPTION OF PURPOSE

CT whole body imaging is widely used for diagnosis, staging and follow up of patients with malignancies. Additionally, bone segmentation, which provides plenty of information to assess bone disease, is used to evaluate metastatic tumor burden in cancer like in prostate or breast cancer.

In prostate cancer, a method assessing this burden on bone scans is already used in clinical practice and is quantified with the Bone scan index. A new method on ⁶⁸Ga-PSMA PET/CT images developed by Bieth et al. [1] showed promising results.

However, in breast cancer no method based on bone scan has shown clinical relevance [2] but literature shows that tumor response can be assessed by evaluating the size and intensity of bone metastasis on PET scans [3]. Our goal is to develop a new method to segment automatically bone and bone lesions on PET/CT images to calculate a new index and prove its clinical relevance to evaluate the tumor response in the context of metastatic breast cancer. The present paper describes the first steps carried out to achieve this goal.

As a previous work showed, the Dice score may not be the best index to evaluate bone segmentation [4], due to its low sensibility to small volume errors, which can be physiologically important. However, this type of errors can be enhanced by the Hausdorff Distance, as it takes into account the largest error in its formulation. This led to think that a Hausdorff Distance-based loss could give better results.

This work compares three approaches to segment the bone structure : one threshold-based and two deep learning based methods with different losses.

2. METHODS

2.1 Data

Sixty patients were recruited in the prospective *EPICURE_{seinmeta}* metastatic breast cancer study (NCT03958136). Imaging data were acquired in two sites using different imaging systems. For this work, we only used the collected CT images. Bone segmentations were generated using the threshold-based method and then manually corrected by 4 non-specialist image processing researchers. For the deep learning approaches, 50 patients, splitted into three groups according to the recruiting site and the extent of skeletal metastatic involvement, were used for training and 10 were used for testing.

2.2 Threshold-based method

The bone structures have specific Hounsfield values allowing the use of thresholds and morphological operations to segment them. The method presented here is inspired by the work of Banik, Rangayyan, and Boag [5]. Figure 1 shows the different steps for bone segmentation. First, the exam table is removed from the original CT to avoid its segmentation as bone. Then, the bone structures are segmented by applying a threshold between 200 HU and 3000 HU, and by carrying out dilation and iterative hole filling tasks. As the bone structure is completely connected by joints, a connected component filter was used to remove small errors found in the heart region as shown in the Figure 1-f. The last step is an erosion operation to prevent over-segmentation.

Further author information: (Send correspondence to Noémie Moreau)
N. Moreau: E-mail: noemie.moreau@keosys.com

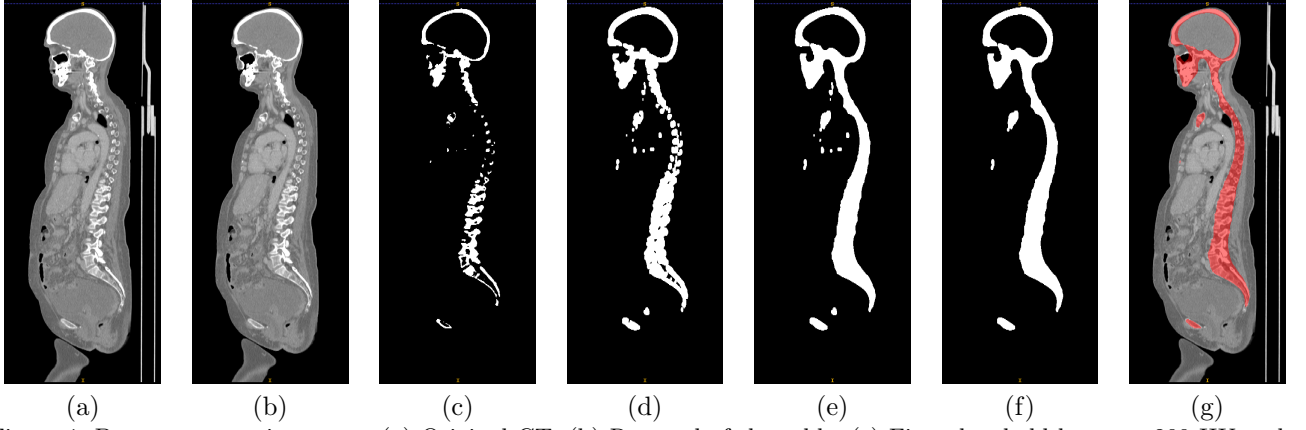


Figure 1. Bone segmentation steps. (a) Original CT, (b) Removal of the table, (c) First threshold between 200 HU and 3000 HU, (d) Dilatation operation, (e) Iterative hole filling, (f) Connected component, (g) Erosion and final result on the CT

2.3 Deep learning methods

2.3.1 Network architecture

The 3D U-Net from the framework “no new net”(nnU-Net) was used in the two deep learning methods. This framework achieved state of the art performance on recent challenge like KiTS2019 [6]. Moreover, it allows to automatically set a number of hyper-parameters given several information, as input data features and memory consumption. It also enables easy integration of new architectures and methods, such as new losses. Except for the losses, all network parameters used in this paper can be found in the original nnU-Net publication [6].

In this work, we compare the performance of the standard 3D nnU-Net with a Cross Entropy/Dice loss, to a tuned 3D nnU-Net using a Hausdorff Distance/Dice loss.

Cross Entropy and Dice loss The sum of the cross entropy loss (L_{CE}) and the dice loss (L_{DCS}) is the loss used in the standard nnU-Net [6]. They are defined by Drozdal et al. [7] as :

$$L_{CE} = \sum_i p_i \log(q_i) + (1 - p_i) \log(1 - q_i)$$

$$L_{DCS} = -\frac{2 \sum_i p_i q_i}{\sum_i p_i + \sum_i q_i}$$

with p the ground truth and q the predicted segmentation.

Hausdorff Distance and Dice loss The Hausdorff Distance is rarely used as loss due to its non derivability. However, Karimi and Salcudean [8] developed three alternative methods to approximate the Hausdorff Distance and use it as loss. Only the one using the Distance Transform (see Figure 2) will be used in this work as it is the closest approximation of the Hausdorff Distance :

$$L_{HD} = \frac{1}{|\Omega|} \sum_{\Omega} ((p - q)^2 \circ (d_p^\alpha + d_q^\alpha))$$

with p the ground truth and q the predicted segmentation, d_p and d_q their respective Distance Transforms and $\alpha = 2.0$ [8].

As the Hausdorff Distance loss focuses only on the largest error which could lead to unstable learning, the loss is balanced with a Dice loss term : $L_{HD+DCS} = L_{DCS} + \lambda L_{HD}$.

In the original paper [8], λ is updated every epoch to give equal weights to each term of the loss : it is the ratio between the mean of L_{HD} and the mean of L_{DCS} over the batch size from the previous epoch. However, experiments showed that the training was still unstable with the automatic parameters set by the nnU-Net. For this reason, λ was fixed to 0 between the 1st and the 250th epoch. Between 250th and the 750th epoch, $\lambda = \text{mean}(L_{DCS})/\text{mean}(L_{HD})/500 * (\text{epoch} - 250)$. After the 750th epoch, $\lambda = \text{mean}(L_{DCS})/\text{mean}(L_{HD})$.

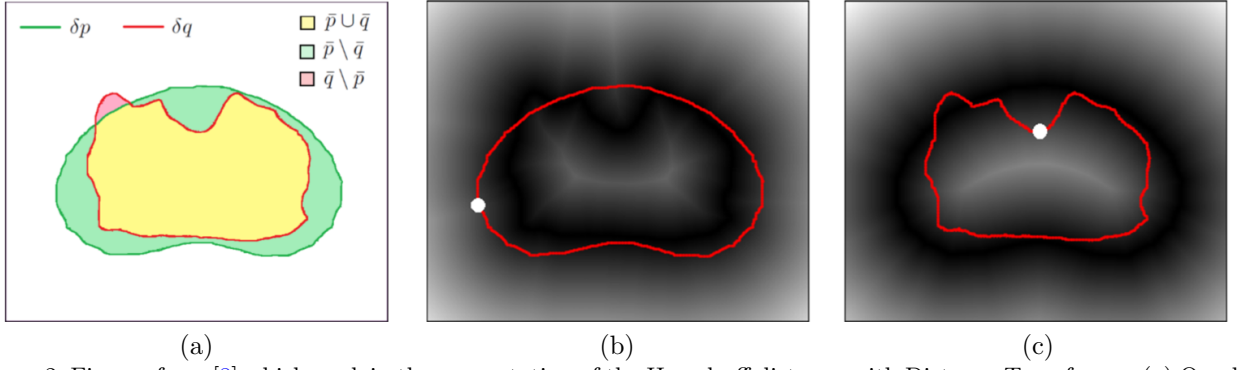


Figure 2. Figures from [8] which explain the computation of the Hausdorff distance with Distance Transforms. (a) Overlap between the ground truth $\bar{p}$ and the predicted segmentation $\bar{q}$. (b) Distance Transform d_q with δp overlaid in red. The white point is the location of $hd(\delta p, \delta q)$. (c) Similar to (b) with Distance Transform d_p and δq overlaid in red, for $hd(\delta q, \delta p)$.

2.4 Evaluation

2.4.1 Quantitative evaluation

Quantitative evaluation was obtained computing the mean recall, mean precision, mean Dice score, mean Hausdorff Distance and the maximum Hausdorff Distance across all test cases. A global Dice score for all test cases combined into one was also performed. For the Dice score and the Hausdorff Distance, a paired t-tests (significance level at $p = 0.05$) was performed to evaluate the significance of the difference between methods.

2.4.2 Qualitative evaluation

Qualitative evaluation was done by visual medical expert evaluation on the 10 testing cases. For all cases, the expert was asked to choose the best segmentation results between the three proposed (one for each method). Segmentations were shown simultaneously on the same CT. The expert could see the CT in the three axes and roam through the slices. He could choose to show all segmentations together or only one or two at the time. Each segmentation was randomly named with a different number in each case. If the difference between two segmentation was too small, the expert could choose both of them.

3. RESULTS

This work compares the results of three bone segmentation methods:

1) A Threshold-based method with threshold and morphological operations. 2) A standard 3D nnU-Net with a Cross entropy/Dice loss ($U-Net_{CE+DCS}$). 3) A tuned 3D nnU-Net with a Hausdorff Distance/Dice loss ($U-Net_{HD+DCS}$).

Results of the quantitative evaluation are shown in table 1. All three methods have a Dice score superior to 0.90 but the Threshold-based method is significantly worse than the two other methods in terms of Dice score ($p - value \leq 0.001$ for both methods) and worse than $U-Net_{HD+DCS}$ method in terms of Hausdorff Distance ($p - value = 0.011$). The two deep learning methods are significantly different in term of Dice score ($p - value \leq 0.001$) but not in terms of Hausdorff distance ($p - value = 0.29$) even if the mean Hausdorff Distance and the maximum Hausdorff Distance are better with the $U-Net_{HD+DCS}$ method.

When conducting the qualitative assessment, the expert never chose the Threshold-based segmentation, as it presents some segmentation errors in the heart and under-segmentation in the vertebrae. He did not observe any significant differences between the two deep learning approaches except for one case where part of the heart was segmented by the $U-Net_{CE+DCS}$ (see Figure 3).

Table 1. Quantitative evaluation

Methods	Mean Recall	Mean Precision	Mean Dice	Global Dice	Mean HD	Max HD
<i>Threshold – based</i>	0.91	0.97	0.94	0.95	82	144
$U-Net_{CE+DCS}$	0.99	0.98	0.99	0.99	53	133
$U-Net_{HD+DCS}$	0.99	0.97	0.98	0.98	41	67

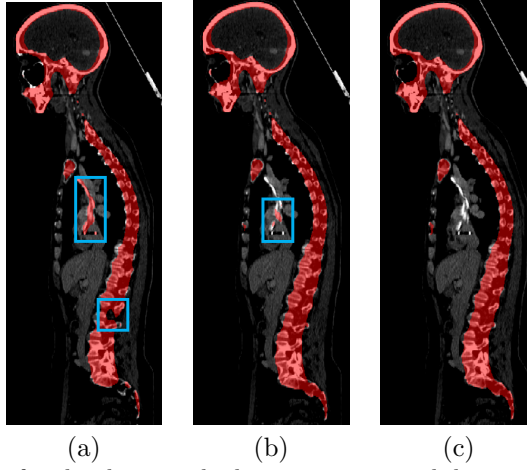


Figure 3. Bone segmentation results for the three methods on a patient with heart calcification. (a) *Threshold – based* : over-segmentation of the heart and under-segmentation of the vertebrae, (b) $U\text{-Net}_{CE+DCS}$: over-segmentation of the heart , (c) $U\text{-Net}_{HD+DCS}$: best segmentation result.

4. NEW OR BREAKTHROUGH WORK

This work compared three approaches to segment the bone in whole-body CT images. All three methods gave good results but the analysis showed that deep learning methods performed best. This work also showed that a network with Hausdorff/Dice-based loss gave better result in term of Hausdorff Distance than a CE/Dice loss. Even if the quantitative results were not significantly different, the visual evaluation for one patient showed that when the $U\text{-Net}_{CE+DCS}$ method over-segment the heart it was not the case with the $U\text{-Net}_{HD+DCS}$ method. Actually, a previous work showed that a network trained with a CE/Dice loss tend to over-segmented part of the heart, when because of calcification it presents the same Hounsfield value as the bone [4]. Preliminary results are encouraging and the two methods should be evaluated with more data to see if they can be generalized.

ACKNOWLEDGMENTS

This work is partially financed through "Programme opérationnel régional FEDERFSE Pays de la Loire 2014-2020" noPL0015129 (EPICURE) and by the SIRIC ILIAD Nantes-Angers-INCADGOS-Inserm 12558 grant.

The $EPICURE_{seinmeta}$ study was approved by the French Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM, 2018-A00959-46) and the Comité de Protection des Personnes (CPP) IDF I, Paris, France (CPPIDF1-2018-ND40-cat.1), and a written informed consent was signed by each participant.

References

- [1] M. Bieth et al. "Exploring New Multimodal Quantitative Imaging Indices for the Assessment of Osseous Tumor Burden in Prostate Cancer Using ^{68}Ga -PSMA PET/CT." In: *Journal of Nuclear Medicine* (2017).
- [2] M. Colombié et al. "Évaluation d'une méthode de quantification de la masse métastatique osseuse par mesure automatisée du Bone Scan Index, dans le suivi thérapeutique des cancers du sein". In: *Médecine Nucléaire* 4101.6 (2013), pp. 233–273.
- [3] Ukihide Tateishi et al. "Bone Metastases in Patients with Metastatic Breast Cancer: Morphologic and Metabolic Monitoring of Response to Systemic Therapy with Integrated PET/CT". In: *Radiology* 247.1 (2008), pp. 189–196.
- [4] N Moreau et al. "Deep learning approaches for bone and bone lesion segmentation on ^{18}F FDG PET/CT imaging in the context of metastatic breast cancer". In: *Engineering in Medicine and Biology Conference* (2020).
- [5] S. Banik, R. Rangayyan, and G. Boag. *Landmarking and Segmentation of 3D CT Images*. Morgan & Claypool, 2009.
- [6] Fabian Isensee et al. "Automated Design of Deep Learning Methods for Biomedical Image Segmentation". In: *arXiv preprint arXiv:1904.08128* (2020).
- [7] M. Drozdal et al. "The importance of skip connections in biomedical image segmentation". In: *Deep Learning and Data Labeling for Medical Applications* (2016), pp. 264–271.
- [8] D. Karimi and S. E. Salcudean. "Reducing the Hausdorff Distance in Medical Image Segmentation With Convolutional Neural Networks". In: *IEEE Transactions on Medical Imaging* 39.2 (2020), pp. 499–513.