EPiC Series in Health Sciences

Volume 7, 2024, Pages 155–159

Proceedings of The 24th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery

SSPC: A new topological metric for deep learning based anatomical reconstruction evaluation.

Lhoussein Axel Mabrouk[1], Fabrice Bertrand[1], Francois Boux de Casson[1], and Clément Daviller[1]

[1]Blue Ortho an Exactech company, Meylan, France

Lhoussein.mabrouk@blue-ortho.com

Abstract

Computer-assisted surgery relies on precise labeling of patient anatomy using 3D images. Major part of this process is nowadays performed by deep-learning (DL) algorithms. However, the evaluation of automated segmentations using conventional metrics like Dice coefficient or Hausdorff distance has limitations, especially when assessing non-significant errors at the mesh level. To overcome this, we propose a novel metric (SSPC) focusing on significant surface disparities to enhance evaluation accuracy.

1 Introduction

In the field of orthopedics, the integration of artificial intelligence (AI) for bone segmentation is very promising. Accurate bony structures modeling is crucial for surgical planning and computer-assisted navigation. The growing AI usage in this context requires robust evaluation metrics for ensuring 3D mesh reliability.

Constructing a 3D bone model through image segmentation requires labeling voxels by distinguishing bone and non-bone regions. The resulted mask is often converted to a mesh for clinical use. The evaluation metric must align with the specific application domain's nuanced requirements [1] [2]. Specifically, in the context of orthopedics, there are two main challenges:

1. Doing assessment at the mesh level, instead of evaluating the mask voxels. This allows a more stable score regardless of image resolution and voxel size [3] [4] [5].
2. Exclusively penalizing errors that bear significant clinical implications, which requires using a precise threshold from a continuous scale [6] to filter surface noise bias.

Given the inability of conventional metrics to address these constraints, we have introduced the Substantial Surface Proximity Classifier(SSPC), that facilitates the clinical modulation of noise tolerance.

2 Limitations of existing metrics

Two widely adopted approaches are generally employed for reconstruction quality assessment:

- Ensemble metrics, directly on the volume voxels. Those are directly or indirectly based on the cardinalities of the confusion matrix [7]. Consequently, inaccuracies at external surface are not highlighted on the ultimate score. Conversely, in cases having non-clinically significant noise, it can detrimentally impact the metric, contradicting its intended use. Among the most used in this category, we find Continuous Dice(CD) [8], IOU [9], MCC [10], AUC and ROC curves [11]
- Alternatively, methodology involves assessing the external surface using metrics such as the Hausdorff Distance (HD) [12]; however, addressing noise with a high degree of selectivity proves challenging. The ASSD [6], VSNR[13] and RMS[14] are other commonly employed metrics; nevertheless, sensitive to outliers.

An additional interesting metric is the Normalized Surface Distance (NSD)[15], affording more flexible noise filtering. However, it was originally designed to process pixels within a mask, posing a challenge once the mask is transformed into a mesh. MSDM2 is also an interesting multi-scaled metric, designed by G.Lavoué [3] to better represent the visual distortion of meshes for general graphic rendering. Yet this latter being insufficient in a clinical context.

3 Substantial Surface Proximity Classifier (SSPC)

3.1 Methodology

SSPC metric consists of measuring cloud-to-mesh distances between two 3D objects in both directions. Using the union of the two clouds $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$, all vertices $\boldsymbol{V_{hT}}$ located beyond a tolerance distance $\boldsymbol{T}$ are counted. This distance must be set according to the targeted clinical application. The remaining $\boldsymbol{V_T}$ vertices are considered insignificant and are excluded.

$$V_{hT} = |\{\mathrm{v} \in C_1 U\ C_2 |\ \mathrm{v} \geq \mathrm{T}\}| \tag{1}$$

The ***SSPC*** *score* is then the ratio between $\boldsymbol{V_T}$ and the total number of vertices.

$$SSPC = \frac{V_T}{V_{hT} + V_T} \tag{2}$$

3.2 Validation process

Reconstructions produced by two DL models, denoted as *M1* and *M2*, on a cohort of 103 cases, were evaluated using the SSPC, CD, HD, and its derivatives (HD95, and HD98). These reconstructions, were compared to a ground truth (GT), generated by 3 experts with over 5 years of experience. The scores were used to identify the most faithful reconstruction among those from *M1* and *M2*. These reconstructions were also presented to the same experts, asking them to select the most consistent result. Their answers were confronted to the metrics outputs to determine which metric was most relevant to human expertise.

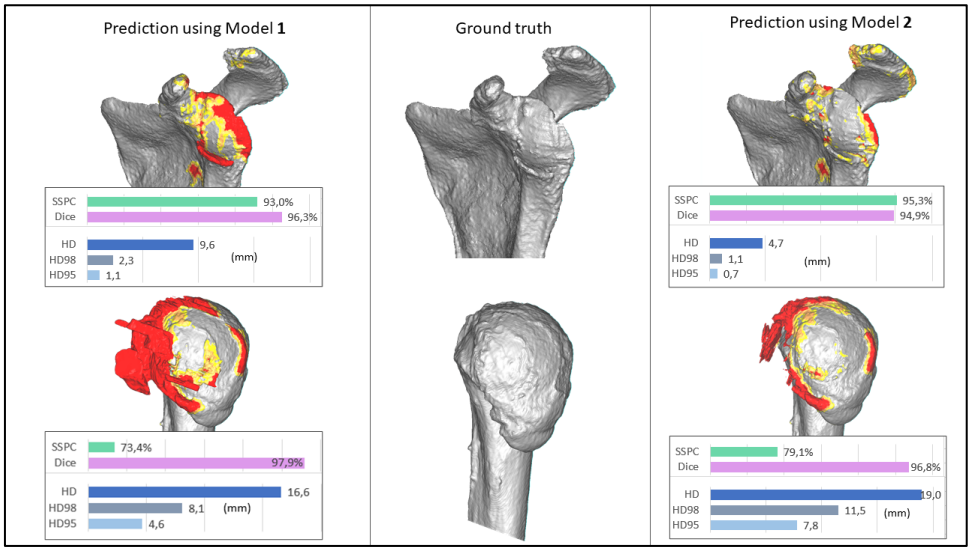
4 Results

For our experiments, we chose an SSPC tolerance $T=0.5$, which corresponds to the voxel size. Thus, the degree of reliability obtained between the tested metrics and the experts, on the 103 evaluated cases, is presented in the Table 1. It was respectively 67.0%, 76.7%, 62.1%, 75.7%, and 87.4 % for the CD, HD, HD98, HD95 and SSPC; which proves the greater relevance of the latter.

	CD [8]	HD [12]	HD98	HD95	SSPC
Expert 1	74,4%	76,9%	79,5%	66,7%	87,2%
Expert 2	61,5%	71,8%	69,2%	56,4%	87,2%
Expert 3	64,0%	84,0%	80,0%	64,0%	88,0%
All Experts	**67,0%**	**76,7%**	**75,7%**	**62,1%**	**87,4%**

Table 1: The degree of reliability of the evaluation metrics compared to human expertise.

In Figure 1, we showcase the scores yielded by the assessed metrics for two bones, reconstructed using distinct models. The reconstruction errors are estimated against the GT in the middle.



It shows that:

Figure 2: Illustration of how SSPC addresses the limits of conventional metrics, exemplified in the assessment of DL derived bone reconstructions. Mesh vertices are classified into four groups by distance from the GT: <0.5mm in gray, <1mm in yellow, <2mm in orange, and >=2mm in red.

1. For the scapula bone case (above), CD[8] is more favorable for *M1*(96.3%), even though it entails more significant errors. However, this score was lower for *M2*(94.9%) due to insignificant hidden noise. Conversely, SSPC scores are in line with the visualized errors, and thus favors the results of *M2*.
2. Regarding the humerus bone depicted at the bottom, the HD[12] and its derivatives advocate the result of *M1*, with respective scores of (16.6, 8.1, and 4.6). In contrast, the M2, which contains fewer errors, is penalized with higher scores (19, 11.5 and 7.8) due to the outliers. This undesired phenomenon is better addressed by SSPC, which favors *M2*(79.1%) over *M1*(73.4%).

5 Conclusion

This cohort study showcased SSPC's superior ability over traditional metrics in identifying significant differences among 3D anatomical objects, as recognized by experts. Unlike conventional metrics [12] [8], it selectively focuses on impactful differences, demonstrating reliable comparison between two reconstruction systems. We are currently exploring methods to integrate SSPC into the model training phase for improved precision.

References

[1] A. Reinke and H. Müller, "Metrics reloaded: a new recommendation framework for biomedical image analysis validation," *Proceedings of the Medical Imaging with Deep Learning (MIDL 2022),* p. 3, 2022.

[2] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging,* p. 29, 2015.

[3] G. Lavoué, "A Multiscale Metric for 3D Mesh Visual Quality Assessment," *Computer Graphics Forum,* vol. 30, no. 5, pp. 1427-1437, 2011.

[4] G. Lavoué, I. Cheng and A. Basu, "Perceptual Quality Metrics for 3D Meshes: Towards an Optimal Multi-attribute Computational Model," in *IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, 2013.

[5] G. Lavoué, M. C. Larabi and L. Váša, "On the Efficiency of Image Metrics for Evaluating the Visual Quality of 3D Models," *IEEE Transactions on Visualization and Computer Graphics,* vol. 22, no. 8, pp. 1987-1999, 2016.

[6] A. Reinke, M. D. Tizabi, C. H. Sudre, M. Eisenmann, T. Rädsch, M. Baumgartner, L. Acion, M. Antonelli, T. Arbel, S. Bakas, P. Bankhead, A. Benis, M. Blaschko, F. Buettner and M. J. Cardoso, "Common Limitations of Image Processing Metrics: A Picture Story," arXiv, 2023.

[7] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," *Mach. Learn. Technol.,* 2008.

[8] R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro and N. Harel, "Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations," bioRxiv, 2018.

[9] N. C. Chung, B. Miasojedow, M. Startek and A. Gambin, "Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data," *BMC Bioinformatics,* p. 644, 2023.

[10] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient ({MCC}) over F1 score and accuracy in binary classification evaluation," *BMC Genomics,* p. 6, 2020.

[11] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters,* pp. 861-874, 2006.

[12] J. Henrikson, "Completeness and Total Boundedness of the Hausdorff Metric," *MIT Undergraduate Journal of Mathematics,* p. 69–80, 1999.

[13] D. M. Chandler and S. S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *IEEE Transactions on Image Processing,* vol. 16, no. 9, pp. 2284 - 2298, 2007.

[14] Wikipedia, "Root-mean-square deviation," [Online]. Available: https://en.wikipedia.org/wiki/Root-mean-square_deviation. [Accessed 29 12 2023].

[15] S. Nikolov, S. Blackwell, A. Zverovitch, R. Mendes, M. Livne, J. D. Fauw, Y. Patel, C. Meyer, H. Askham, B. Romera-Paredes, C. Kelly, A. Karthikesalingam, C. Chu and Carn, "Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study," *Journal of Medical Internet Research,* pp. 9-10, 2021.