EPiC Series in Health Sciences
Volume 7, 2024, Pages 53–57
Proceedings of The 24th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery

Feature-based training and evaluation of deep learning models for shoulder bone 3D reconstruction

Clément Daviller[1], François Boux de Casson[1], Fabrice Bertrand[1], and Lhoussein Axel Mabrouk[1]
[1]Blue Ortho an Exactech company, Meylan, France
clement.daviller@blue-ortho.com

Abstract

Introduction of deep-learning (DL) in the computer assisted surgery requires to train and evaluate segmentation models by maximizing the control and knowledge of data. In this study, we highlight the incomes of data mastery through the examples of shoulder bone segmentations.

1 Introduction

During the last decade, the use of numerical 3D scene reconstruction from CT images for total shoulder arthroplasty in computer-assisted guidance has dramatically grown[1]. This soar comes with the need of accurate segmentation tools, frequently relying on efficient DL methods[2]–[5]. Human health being at stake, the results often require a systematic expert verification and correction to guarantee accuracy. This is especially true with unusual pathological cases, unexpected image acquisition, revision cases or presence of metal artifacts.

Various factors including network architecture, fine tuning but most importantly, the quality of training data can make difference to end with a successful segmentation.

This study seeks to demonstrate the relevance of data mastery on the prediction accuracy through the examples of scapula and humerus segmentation.

2 Material and method

Two datasets have been used for the training of models, dedicated for scapula and humerus. These datasets are composed of 1209 and 1098 reconstructions, built by experienced and qualified operators. The cases were meticulously chosen according to their features, so that the models can be trained with features equally represented. Considered features were gender, shoulder side, fusion between the scapula and humerus, presence of noise, contrast agent, osteophytes, fractures, presence of implant and in this particular class, the type of implants.

Model fitting was done by establishing training, validation and test datasets respecting the above-mentioned feature balancing rules, and by iterative hyperparameters exploration.

Evaluation was made by comparing model predictions to the original ones, reviewed and corrected by an expert, and that was exploited for surgery. The comparison was established, based on classical Dice coefficient and Hausdorff distance with elimination of 5% outliers, and also using a 3D meshes surface similarity-based metric, since it is these latter that are used during surgery. This home-designed metric is adapted from surface Dice similarity coefficient (SDSC) presented in [6] to work on 3D meshes. To perform the comparison, we computed the distances between each vertex of the prediction and the reference mesh. Then, the vertices are labeled as follows:

- negligible error (NE) when distance ≤0.5mm
- small error (SE) when distance ranging between 0.5mm and 1mm
- moderate error (ME) when distance ranging between 1mm and 2mm
- important error (IE) when distance >2mm

3 Results

Table-1 presents the accuracy results, measured in the subgroups comprising the datasets. The overall NE rate for scapula is 96.60% but drops to 94.56% with presence of metal artifact. Conversely, it reaches 98.75% on standard cases (p-value≤0.001). Regarding the humerus, the overall NE rate is 93.07%. It falls to 86.65% in presence of metal artifact and increases to 94.48% with standard cases (p-value≤0.001).

Significant difference was found between standard cases (NE=94.48%) and cases with fusion in humerus (NE=92.87%) but not for the scapula.

Dice coefficient agreed with SDSC as scapula met a median score of 97.66% for standard cases and 95.87% for metal-artefact (p-value≤0.001). Humerus bone met a median Dice score of 98.85%% for standard cases and 97.83% for metal-artefact cases (p-value≤0.01).

For both scapula and humerus, HD95 was not significant as p-value was equal to 0.39 and 0.07 respectively.

Figure 1 gives a visual estimate of these scores with a representation of the error against the mesh reconstructions.

	Feature	Nb cases	Mesh SDSC				Dice coefficient (%)		Hausdorff distance 95%(mm)	
			NE rate (%)	SE rate (%)	ME rate (%)	IE rate (%)	Median	min/ max	median	min/ max
Scapula	**All cases**	150	96.60	98.23	99.03	0.97	96.73	82.52 / 98.56	0.50	0.50/ 33.82
	Standard Cases	44	98.75	99.20	99.45	0.55	97.66	93.68 / 98.56	0.50	0.50/ 2.00
	Noisy Cases	12	98.13	99.41	99.84	0.16	97.44	94.59 / 98.02	0.50	0.50/ 0.71
	Cases with implant or metal artifact	75	94.56	97.15	98.47	1.53	95.87	82.52 / 98.14	0.50	0.50/ 33.82
	Fusion	19	98.76	99.47	99.77	0.23	97.40	94.71 / 98.04	0.50	0.50/ 0.50
Humerus	**All cases**	95	93.07	96.01	97.66	2.34	98.55	84.46 / 99.34	0.5	0.50/ 17.72
	Standard Cases	60	94.48	96.81	98.00	2.00	98.85	84.46 / 99.33	0.5	0.50/ 17.72
	Noisy Cases	6	94.74	97.93	99.51	0.49	98.58	98.26 / 99.34	0.71	0.50/ 1.12
	Cases with implant or metal artifact	14	86.65	91.72	95.21	4.79	97.66	91.73 / 98.43	1.81	0.71/ 15.12
	Fusion	19	92.87	96.15	97.97	2.03	98.49	95.84 / 99.34	0.5	0.50/ 7.79

Table-1: scapula and humerus segmentation accuracy based on mesh SDSC, Dice coefficient and Hausdorff distance 95%.

4 Conclusion

Subgroup analysis allowed accurate assessment of segmentation model performances. Noise had little impact on accuracy. Bone fusion only impacted on humerus segmentation, suggesting that scapula mask should be preferred in overlapping scenarios. Most of all, the presence of metal implant importantly affects the segmentation accuracy either for scapula or humerus.

Given the unpredictable nature of neural network model from input signal, a fine mastery of the features allows deep insight into DL-based algorithms and a better knowledge of their accuracy potential.

Moreover, it also suggests that a well balancing of feature representation in data leads to better performances of AI model. As DL-models trend to be overconfident when making predictions[7], [8] especially with unseen data, inclusion of prior feature might be a key-factor for assessment of their reliability.

Despite numerous metrics and datasets dedicated for image segmentation model accuracy evaluation[9], the assessment by feature is rarely mentioned in literature. Though, it allows a fine rating

of models and identification of their limitations. This lack of description in literature may be due to the fact that this approach requires large and various datasets without what it becomes hard to reach significant conclusions.

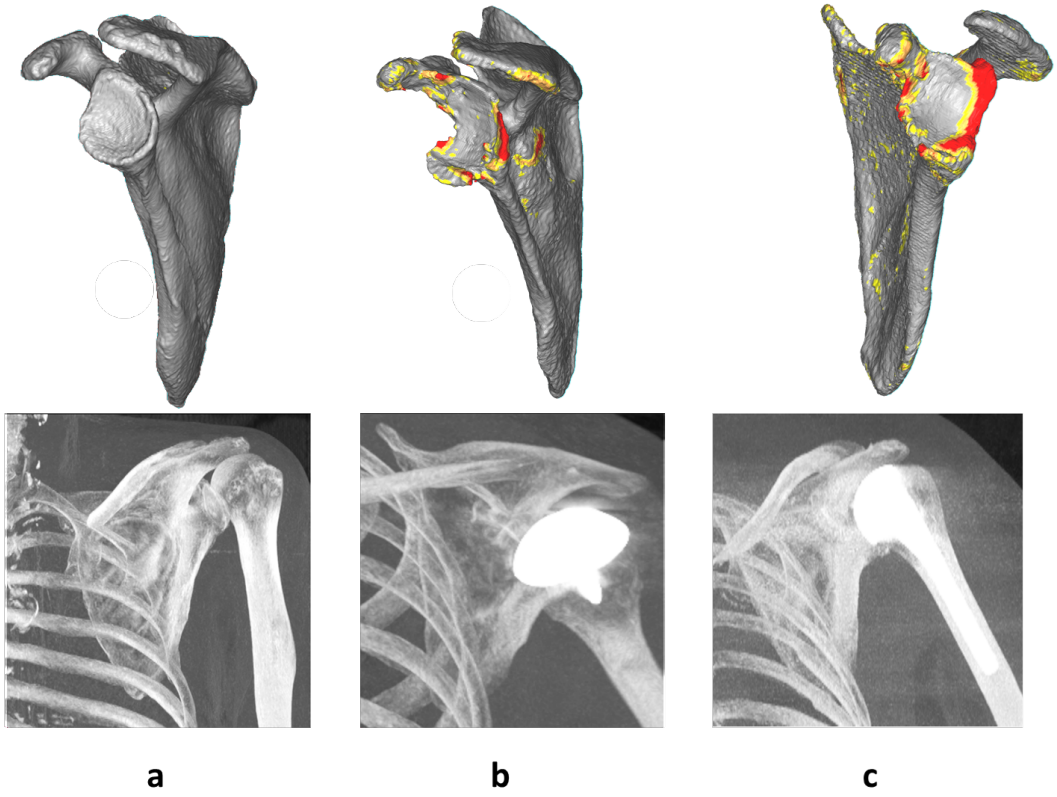

Figure-1: segmentations of 3 different cases (a)without implant (b)with a stemless and (c)with a stem implant. Gray vertices carry a negligible error (NE) in comparison to the reference mesh. Yellow, orange, and red represent respectively vertices carrying small, moderate, and important error.

5 References

[1] X. Fan, Q. Zhu, P. Tu, L. Joskowicz, and X. Chen, "A review of advances in image-guided orthopedic surgery," *Phys. Med. Biol.*, vol. 68, no. 2, p. 02TR01, Jan. 2023, doi: 10.1088/1361-6560/acaae9.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597 [cs]*, May 2015, Accessed: Mar. 22, 2021. [Online]. Available: http://arxiv.org/abs/1505.04597

[3] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," arXiv.org. Accessed: Dec. 15, 2023. [Online]. Available: https://arxiv.org/abs/1606.04797v1

[4] G. Schmitt, A. Greene, S. Polakovic, N. Davis, and F. Bertrand, "Results of a Machine Learning Algorithm for Automatic Three-Dimensional Segmentation of Computed Tomography Scans of the Shoulder," *ORS conference*, 2021.

[5] J. Chen *et al.*, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation." arXiv, Feb. 08, 2021. doi: 10.48550/arXiv.2102.04306.

[6] S. Nikolov *et al.*, "Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study," *J Med Internet Res*, vol. 23, no. 7, p. e26151, Jul. 2021, doi: 10.2196/26151.

[7] L. Huang, S. Ruan, and T. Denœux, "Application of belief functions to medical image segmentation: A review," *Information Fusion*, vol. 91, pp. 737–756, Mar. 2023, doi: 10.1016/j.inffus.2022.11.008.

[8] K. Zou, Z. Chen, X. Yuan, X. Shen, M. Wang, and H. Fu, "A review of uncertainty estimation and its application in medical imaging," *Meta-Radiology*, vol. 1, no. 1, p. 100003, Jun. 2023, doi: 10.1016/j.metrad.2023.100003.

[9] A. Reinke *et al.*, "Common Limitations of Image Processing Metrics: A Picture Story." arXiv, Dec. 06, 2023. doi: 10.48550/arXiv.2104.05642.