EPiC Series in Health Sciences
Volume 7, 2024, Pages 25–30
Proceedings of The 24th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery

Automatically Extracted Metrics for Diagnosis of Developmental Dysplasia of the Hip are Sensitive to Assumptions About Morphological Priors

María José Bontá Suárez[1], Emily Schaeffer[2], Kishore Mulpuri[2], Rafeef Garbi[1], Antony J. Hodgson[1]
[1] University of British Columbia, Vancouver, BC, Canada
[2] Orthopedic Surgery, British Columbia Children's Hospital, Vancouver, BC, Canada
mbontas@student.ubc.ca

Abstract

Deep learning techniques for diagnosing Developmental Dysplasia of the Hip (DDH) in newborns from ultrasound (US) images of the hip have demonstrated improved reliability over manual annotations of US scans. While volumetric 3D US has been shown to better represent hip bone morphology, most of the proposed automatic diagnostic approaches to measure 3D equivalents of the commonly used 2D Graf angle rely on strong morphological (geometric) priors. We have found that a significant fraction of cases (~20%) result in metrics which expert assessors regard as incorrect or implausible. We hypothesize that the lack of robustness of existing algorithms is due to their assumption that selected morphological priors are always valid, and this may not hold in a number of cases. In this study, we evaluate the differences between extracted DDH metrics based on expert labels and automatic segmentations. We show that a metric extraction process that uses morphological priors is sensitive to relatively small variations in the segmentation results.

1 Introduction

Developmental Dysplasia of the Hip (DDH) refers to anatomical malformations of the hip joint ranging from mild dysplasia to full dislocation. Early diagnosis has been shown to improve prognosis and mitigate the long-term impact in the quality of life of the affected individuals [1, 2], while late presentations and missed diagnoses are estimated to be responsible for ~20%-40% of osteoarthritis in

young adult patients [3]. 2D ultrasound (US) of the hip in the coronal position is widely used to diagnose DDH, but it suffers from low specificity, especially in the hands of novice radiologists [4]. Comon DDH diagnostic metrics extracted from 2D US include the Graf alpha angle (α) [5] and the femoral head coverage (FHC) ratio [6]. Quader (2021) showed that 3D US can better represent the 3D morphology of the hip and that automatic diagnosis can significantly reduce inter-rater variability [7].

The first intrinsically volumetric DDH metrics such as the 3D α angle were based on hip bone voxel labeling from 3D US volumes of the hip, followed by extraction of geometric properties from these labels [8, 9, 10, 11]. In recent work by El-Hariri (2020), the author shows that using deep learning to segment 3D US volumes to obtain predictions for the pelvis and femoral head and a new method to extract 3D equivalents of the α and FHC from these segmentation predictions improved inter-scan repeatability, but the extraction step assumed certain morphological priors such as bounds on the relative location of the femoral head and acetabulum. At this point, it is unclear how sensitive the extracted dysplasia metrics are to differences in the segmentation of anatomical structures versus differences in the metric extraction process. In this study, therefore, we take the first step towards assessing the sensitivity of the calculated DDH metrics to the performance of the deep learning segmentation models. We qualitatively and quantitatively compare the α_{3D} and FHC_{3D} metrics calculated using El-Hariri's (2020) metric extraction method based on (1) expert segmentations of the pelvis and femoral head and (2) segmentations obtained using El-Hariri's deep-learning-based segmentations.

2 Methods

We generate pelvis and femoral head segmentation predictions from an expert-labelled test set database of previously acquired 3D US volumes using trained 3D UNet models, as described by El-Hariri (2021) [12]. We then extract and compare the 3D alpha angles (α_{3D}) and 3D FHC (FHC_{3D}) from both segmentation predictions and expert labels using the method presented by El-Hariri (2020) [11].

2.1 Dataset

Our dataset consists of 115 expert-labeled 3D US volumes of the hip from 34 newborns (average age 7 weeks, 2-15 weeks) acquired at BC Children's Hospital under ethics approvals H14-01448, H18-00131, and H18-02024. The volumes were collected using the SonixTouch Q+ machine and a 4DL14-5/38 Linear 4D probe at its default penetration settings (7.5MHz with an image depth of 4cm). The mechanically swept B-mode slices acquired over an angle of 30° were reconstructed into 38mm^3 volumes. The volumes were then annotated by a graduate student under the supervision and review of our collaborating orthopaedic surgeon to obtain pelvis and femoral head expert labels. Both healthy and dysplastic infant hips were included in this dataset.

2.2 Bone Segmentation

We trained two 3D UNet networks for the tasks of pelvis and femoral head segmentation following El-Hariri's (2021) guidelines on a subset of 64 US volumes from 20 participants. The volumes were down-sampled to 128×128×128 voxels and, given the small data size available, we applied random classical augmentations (such as scaling, rotation, flipping, etc.). In both tasks we optimized the combined BCE Loss and Dice Loss. We evaluated the trained models on the separate test set of 51 volumes from 14 participants to generate pelvis and femoral head segmentation predictions - see Figure 1B.

2.3 DDH Metric Extraction

The metric extraction algorithm presented by El-Hariri (2020) first identifies a region of interest (ROI) composed of the intersection of two voxel spaces ($ROI = ROI_{AP} \cap ROI_S$) defined by (1) the voxels within the slices in the anteroposterior plane within 3.5mm from the center of mass (COM) of the femoral head (ROI_{AP}), and (2) the voxels within a sphere of 15mm radius centered around the femoral head (ROI_S) - see Figure 1A. Within this ROI, the point of maximum Gaussian curvature is found, and used to define the voxels from the straight portion of the ilium (I) and acetabulum (A). Planes are fitted to both I and A, and the angle between the normal to the planes is calculated to obtain α_{3D}. Finally, FHC_{3D} is calculated as the proportion of the femoral head points medial to the plane of I relative to the total points assigned to the femoral head.

2.4 Evaluation Metrics

For each volume in the test set we calculated α_{3D} and FHC_{3D} from both the expert labels and the segmentation predictions and compared them. We also calculated the Dice coefficient between expert labels and segmentation predictions for the pelvis and femoral head - see Figure 1C. To quantify the correlation between segmentation performance and DDH metric variability, we calculated the Pearson correlation coefficient between the extracted DDH metrics and the calculated Dice values. Finally, for qualitative evaluation, we identified examples of coronal slices from US volumes with large α_{3D} differences between segmentation predictions and expert labels.

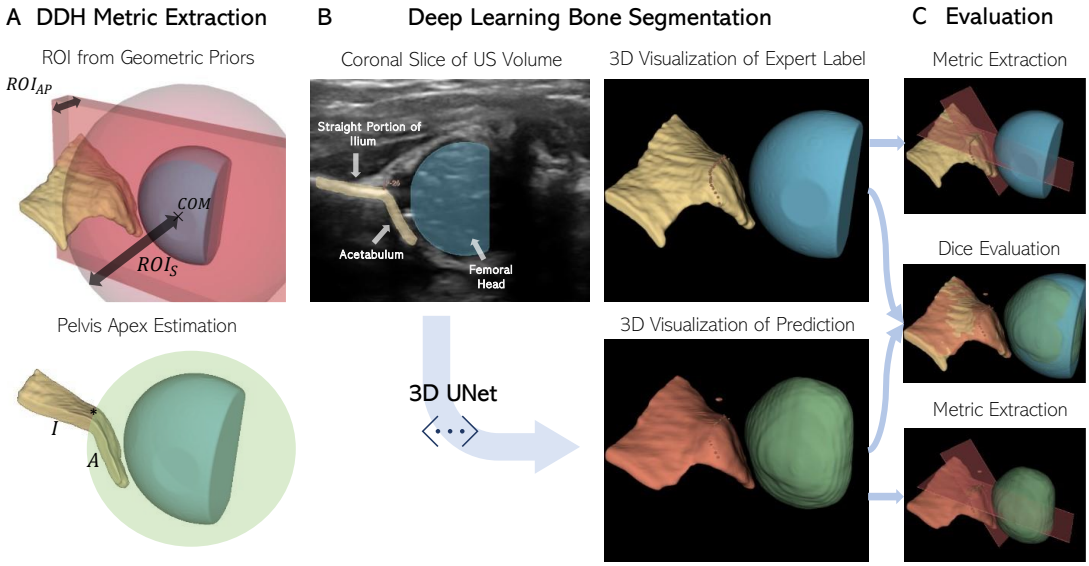

Figure 1. Overview of DDH diagnosis and evaluation framework. **A**. Region of interest (ROI) defined from morphological priors involving the relative position of the femoral head with respect to the pelvis used by the metric extraction approach proposed by El-Hariri (2020) [11]. The points from the pelvis inside a sphere centered at COM with a radius extending to the point of maximum gaussian curvature (*) are assigned to A while the points outside the sphere are assigned to I. **B**. Example of coronal US slice with expert bone segmentations along with a corresponding 3D visualization and deep learning segmentation prediction. **C**. Example of Dice evaluation and plane estimates used for measuring α_{3D} and FHC_{3D}.

3 Results

Figure 2A shows the difference in α_{3D} and FHC_{3D} obtained when comparing metrics extracted from deep learning segmentations and expert labels. The mean differences in α_{3D} and FHC_{3D} using El-Hariri's (2020) metric extraction were -2.0° (SD: 9.8°) and -1.9% (SD: 13.1%) respectively (N=48, 3 samples failed to be calculated). The Pearson correlation coefficients between the absolute α_{3D} difference and the pelvis and femoral head Dice were significant at $\rho = -0.45$ ($p = 0.0020$) and $\rho = -0.341$ ($p = 0.019$) respectively, while the Pearson correlation coefficients between the absolute FHC_{3D} difference and the pelvis and femoral head Dice were not significant at $\rho = -0.244$ ($p = 0.098$) and $\rho = -0.049$ ($p = 0.74$) respectively.

Figure 2B shows examples of 2D coronal slices of volumes with discrepancies in α_{3D} greater than 10°.

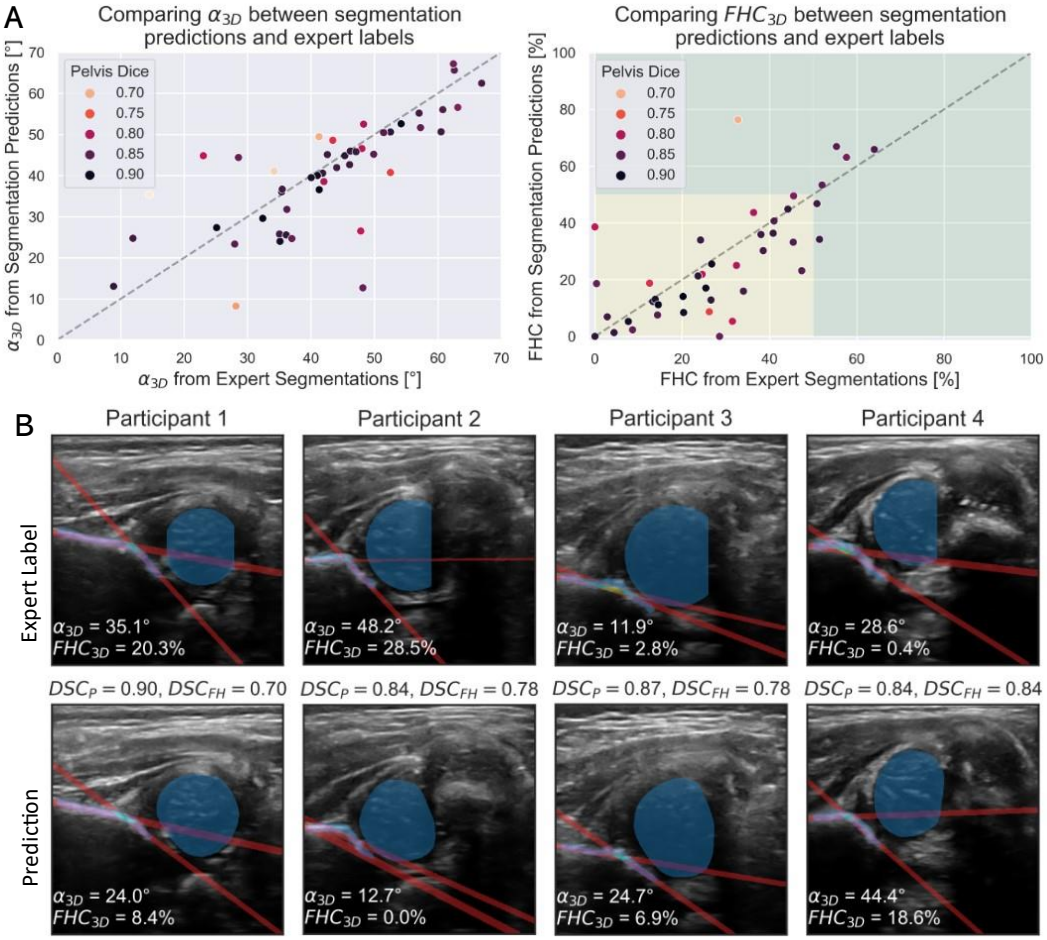

Figure 2. A. Comparison of α_{3D} and FHC_{3D} obtained from expert segmentations and deep learning predictions where the hue indicates pelvis Dice. The black dotted line shows $y = x$. (Right) The background color change in the FHC_{3D} plot indicates the delimitation between a healthy hip ($FHC > 50\%$) against a potentially unstable hip ($FHC < 50\%$) [6]. **B**. Coronal slices of US volumes from 4 different participants with large discrepancies in α_{3D} values between the two segmentation approaches (expert vs deep-learning-based) showing pelvis and femoral head

masks with the extracted I and A planes and calculated α_{3D} and FHC_{3D}. We also show the pelvis Dice (DSC_P) and the femoral head Dice (DSC_{FH}) values.

4 Discussion & Conclusion

We evaluated El-Hariri's (2020) [11] automatic DDH diagnosis method on 3D US volumes of newborn's hips and found that, for a given US volume, there were significant discrepancies between the DDH metrics calculated from expert labels and deep learning bone segmentations, leading to a diagnostic category change in 11/48 (23%) hips based on the Graf classification scheme [5]. Some of the discrepancies in the calculated metrics can be attributed to significant differences between the manual and automatic segmentations of the pelvis and femoral head (e.g., as indicated by low Dice values), but there remains a considerable number of cases where the discrepancy in α_{3D} is above 10° even though the Dice score is well above 0.80.

Although α_{3D} is indicative of pelvis geometry, the small but significant correlation of this value with femoral head Dice presumably arises from the reliance on morphological priors used to find the apex of the pelvis - see Figure 1A. This reliance on morphological priors appears to make the metric extraction process sensitive to the segmentations. This study is the first to note that, even in situations where the test-retest repeatability of a deep-learning-based algorithm appears to be quite good, the deviations between the expert labels and deep-learning segmentations can produce significant discrepancies in the resulting metrics [8, 9, 10, 11]. In future, we intend to develop methods to automatically identify the apex of the hip from 3D US in a manner that does not rely on use of morphological priors in hopes of improving robustness to variations commonly encountered in the clinical setting, which would make these automatic metric extraction algorithms more suitable for clinical deployment.

References

[1] T. Woodacre, A. Dhadwal, T. Ball, C. Edwards and P. J. Cox, "The costs of late detection of developmental dysplasia of the hip," *https://doi.org/10.1007/s11832-014-0599-7*, vol. 8, no. 4, pp. 325-332, 8 2014.

[2] E. K. Schaeffer and K. Mulpuri, "Developmental dysplasia of the hip: addressing evidence gaps with a multicentre prospective international study," *The Medical journal of Australia*, vol. 208, no. 8, pp. 359-364, 5 2018.

[3] L. Gala, J. C. Clohisy and P. E. Beaulé, "Hip Dysplasia in the Young Adult," *The Journal of bone and joint surgery. American volume*, vol. 98, no. 1, pp. 63-73, 1 2016.

[4] E. Mostofi, B. Chahal, D. Zonoobi, A. Hareendranathan, K. P. Roshandeh, S. K. Dulai and J. L. Jaremko, "Reliability of 2D and 3D ultrasound for infant hip dysplasia in the hands of novice users," *European Radiology*, vol. 29, no. 3, pp. 1489-1495, 3 2019.

[5] R. Graf, "Fundamentals of Sonographic Diagnosis of Infant Hip Dysplasia," *Journal of Pediatric Orthopaedics*, vol. 4, no. 6, pp. 735-740, 11 1984.

[6] C. Morin, H. T. Harcke and G. D. MacEwen, "The infant hip: real-time US assessment of acetabular development," *Radiology*, vol. 157, no. 3, pp. 673-677, 1985.

[7] N. Quader, A. J. Hodgson, K. Mulpuri, A. Cooper and R. Garbi, "3-D Ultrasound Imaging Reliability of Measuring Dysplasia Metrics in Infants," *Ultrasound in Medicine & Biology*, vol. 47, no. 1, pp. 139-153, 1 2021.

[8] A. R. Hareendranathan, M. Mabee, K. Punithakumar, M. Noga and J. L. Jaremko, "A technique for semiautomatic segmentation of echogenic structures in 3D ultrasound, applied to infant hip dysplasia," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 1, pp. 31-42, 1 2016.

[9] M. G. Mabee, A. R. Hareendranathan, R. B. Thompson, S. Dulai and J. L. Jaremko, "An index for diagnosing infant hip dysplasia using 3-D ultrasound: the acetabular contact angle," *Pediatric Radiology*, vol. 46, no. 7, pp. 1023-1031, 6 2016.

[10] N. Quader, A. Hodgson, K. Mulpuri, A. Cooper and R. Abugharbieh, "Towards reliable automatic characterization of neonatal hip dysplasia from 3D ultrasound images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9900 LNCS, pp. 602-609, 2016.

[11] H. El-Hariri, "Reliable and robust hip dysplasia measurement with three-dimensional ultrasound and convolutional neural networks," 2020.

[12] H. El-Hariri, A. J. Hodgson, K. Mulpuri and R. Garbi, "Automatically Delineating Key Anatomy in 3-D Ultrasound Volumes for Hip Dysplasia Screening," *Ultrasound in Medicine & Biology*, vol. 47, no. 9, pp. 2713-2722, 9 2021.