

INFANT HIP SCREENING USING MULTI-CLASS ULTRASOUND SCAN SEGMENTATION

Andrew Stamper¹ Abhinav Singh^{1,2} James McCouat^{1,2} Irina Voiculescu¹

¹ Department of Computer Science, University of Oxford, UK

² NDORMS, University of Oxford, UK.

ABSTRACT

Developmental dysplasia of the hip (DDH) is a condition in infants where the femoral head is incorrectly located in the hip joint. We propose a deep learning algorithm for segmenting key structures within ultrasound images, employing this to calculate Femoral Head Coverage (FHC) and provide a screening diagnosis for DDH. To our knowledge, this is the first study to automate FHC calculation for DDH screening. Our algorithm outperforms the international state of the art, agreeing with expert clinicians on 89.8% of our test images.

Index Terms— Semantic Segmentation, DDH ultrasound

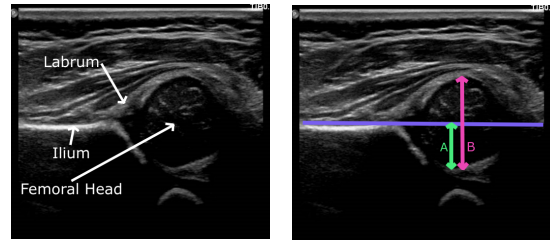
1. INTRODUCTION

Developmental dysplasia of the hip (DDH) is a common cause of childhood disability, with missed cases representing the largest cause of premature arthritis in young adults [1]. In newborns, the condition ranges from outward migration of the femoral head to complete dislocation from the socket. DDH diagnosis under three months enables non-invasive treatment; whereas late diagnosis necessitates expensive complex surgery with significant interruption to social development in infancy [1, 2]. Approximately one third of hip replacements in patients under the age of 40 are due to DDH [1].

Employing experts to perform and report scans routinely is a significant cost barrier to scanning *all* newborns. Therefore, only those with risk factors undergo an ultrasound.

Ultrasound imaging is more sensitive (clinically) than a physical examination and is ideal for screening programmes as it is safe and portable [3]. The Femoral Head Coverage (FHC) method is used to provide objective evidence where a straight line is drawn extending along the upper edge of the ilium (Fig. 1) and the femoral head proportion on each side of the line is measured [4]. Clinically, a decision value of 50% coverage is used, with a larger percentage below the line ($FHC > 50\%$) identified as healthy and a larger percentage above the line ($FHC \leq 50\%$) termed DDH [5].

State of the art literature for automatic DDH evaluation uses the Graf method [6, 7, 8]. However, this method has extremely high inter-operator variability and it only achieves approximately 85% agreement with clinicians. Our reproducible algorithm takes automated numerical measures of the



(a) Annotated hip joint (b) FHC Measurements

Fig. 1. (a) Static ultrasound of hip joint, annotated with femoral head, ilium and labrum. (b) Measurements required for FHC evaluation: $FHC\% = \frac{A}{B} \times 100$.

FHC (AFHC) to classify the presence or absence of DDH. It does so by segmenting key anatomical structures from static 2D ultrasound images: ilium, femoral head and labrum (Fig. 1). Clinicians are able to identify these structures and diagnose by simply inspecting the scan (Gestalt laws) [9].

In this study, two experienced clinicians have generated the one-bit ground truth determining the presence or absence of DDH as perceived through their clinical experience. Hereafter, this clinical decision is referred to as the Gestalt Ground Truth (GGT). Separately, a third clinician has segmented out anatomical features and has taken FHC measurements, referred to as the Segmented Ground Truth (SGT).

2. METHOD

Our algorithm first uses a multi-class convolutional neural network to segment the key anatomical structures mentioned earlier (Fig. 2b). The presence of these structures is a marker that the correct ultrasound image has been acquired. To measure AFHC, the algorithm uses the raw segmentation results to locate the extrema on the femoral head, and the straight line extending from the uppermost edge of the ilium. Our algorithm locates the upper and lowermost points of the femoral head by first locating the centroid of the head mask and then flooding the mask (Fig. 2c).

To identify the ilium line (Figure 2d) we first isolate its horizontal part. We identify the topmost pixel of the ilium mask (x, y) s.t. $(x, y) \in \text{mask}$ but $(x, y+1) \notin \text{mask}$, then parti-

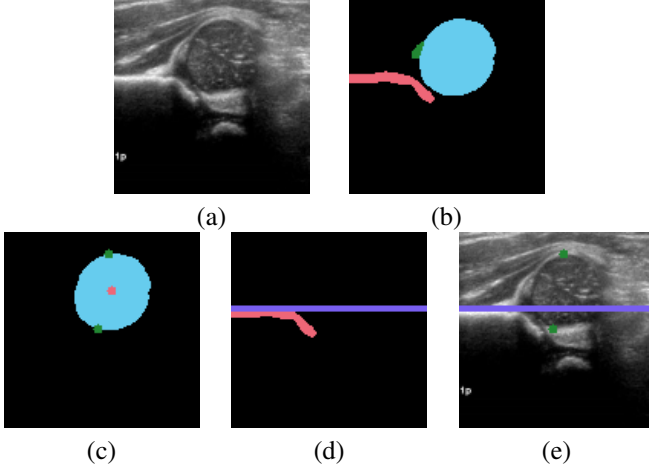


Fig. 2. Generating AFHC. (a) Raw image, (b) Automated segmentation, (c) Femoral head centroid (red) and extrema (green), (d) Ilium line (purple), (e) Key points for AFHC. Inferred is the diagnosis that DDH is present.

tion the boundary into short segments. A line of best fit is drawn through each segment, and the angles between adjacent lines are evaluated. The point at which the ilium turns (downwards) from horizontal is therefore the point at which this angle is maximal. Pixels to the left of this point are identified as part of the horizontal upper ilium edge and a horizontal line is drawn through their average height (Fig. 2e)

Using the two extrema of the femoral head, and the produced ilium line (Fig. 2e) determine AFHC, and thus DDH diagnosis is derived. A decision value of $FHC \leq 50\%$ is used to provide a screening diagnosis for DDH.

A lightweight multi-class U-Net CNN is used to segment the key anatomical structures. This U-Net is a variant of that introduced in [10], modified to use 128×128 input and output, down sampling to a resolution of only 4×4 pixels, and ‘SAME’ padding. Our U-Net encoder comprises six ‘units’, each built from: a 3×3 convolution, batch normalisation and ReLu function. Between each unit of the encoder 2×2 Max Pooling is applied. The decoder is built symmetrically, replacing Max Pooling with 2×2 Up Convolutions. A batch size of 4 and an ADAM optimizer (learning rate: 0.001) are used with the cross entropy loss function. Data augmentation reduces the risk of over-fitting. The augmentations are conducted in a random order, with random degree within the following ranges. Overall image brightness: 0.5–1.5, Gaussian blur: sigma 0.0–2.0, x -axis or y -axis scaling: 0.9–1.1, x -axis or y -axis translation: 0–10%, rotation: -15° to 15° .

The model was built in TensorFlow, and trained for 50 epochs. When compared with the standard U-Net, the reduced number of parameters makes our model particularly lightweight, taking less than 15 seconds per epoch to train on a low power laptop CPU: Intel i7-8550U@1.80GHz.

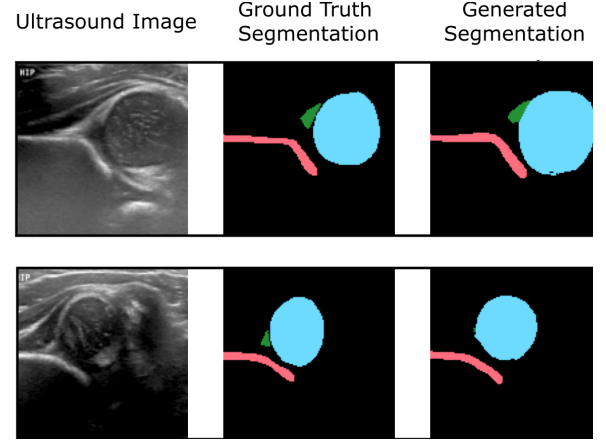


Fig. 3. Qualitative display of automated segmentation results.

3. DATASET

The data has been obtained directly from static ultrasounds taken at the Alder Hey Children’s Hospital in Liverpool, England as part of routine screening. Scanning was performed using the Philips EPIQ5G (L12-5 linear probe) ultrasound machine (Philips, Amsterdam, Netherlands). All images were taken at under 12 weeks and each subject had either an abnormal clinical examination for DDH, or a positive risk factor for DDH [11].

To generate the GGT, two experts classified the images into three groups: Normal (centred), and two severities of DDH: Dysplastic or Dislocated (requires monitoring and treatment). A third clinician derived the SGT by segmenting three key structures from each ultrasound scan: the femoral head, ilium and labrum (Fig. 2b). These masks were verified by the two experts.

The dataset contains 94 right and 96 left hip scans. The left hip scans are reflected to produce 190 right-hip-like scans. Data was split into testing, validation and training by a 50:25:25 % split. To ensure a fair representation of each type of scan in each of these sets, the dataset was first split into normal (71 images, 37%), dysplastic (66 images, 35%) and dislocated collections (53 images, 28%) (according to GGT), and these collections were combined proportionally to create balanced training, validation and testing sets.

The raw images were of non uniform width and height. Each image was pre-processed, cropping to 384×384 pixels, then down sampled (via max pooling) to 128×128 pixels. No further image harmonisation was completed.

4. RESULTS AND DISCUSSION

The segmentation model identifies all three anatomical structures in all scans in the test set (Tab. 1). The femoral head has good overlap scores (Dice, Precision, Recall, Sensitivity, Specificity, etc.) but, as expected from the presentation of the

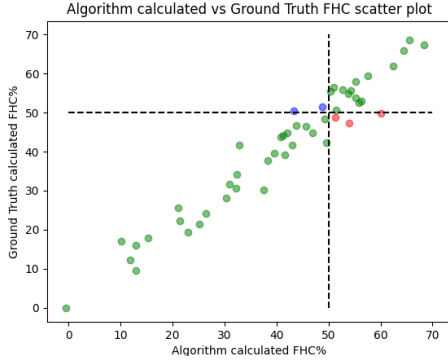


Fig. 4. Scatter plot of AFHC diagnosis versus SGT. The 50% decision value is shown in black. Patients with correct diagnosis in green, false positives in blue, false negatives in red.

data, less accurate boundaries (Hausdorff and Average Symmetric Surface Distance). The ilium is well identified in both overlap and Average Symmetric Surface Distance scores but performs less impressively in Hausdorff Distance, implying the majority of boundary pixels are well identified but the odd pixel in the boundary may be displaced (since HD is the maximal boundary displacement across the whole image). While overlap scores of the labrum appear somewhat unimpressive, they serve well the purpose of its identification.

Fig. 3 compares the SGT and automated segmentation of two sample images: one typical performance and one of its worst segmentations (identifying the labrum by only five pixels). Having compared the segmented masks against SGT segmentation masks, we focus on our primary interest in the system’s diagnostic performance. Comparing the diagnosis inferred from AFHC against SGT we obtain 89.8% accuracy, 90.6% sensitivity and 88.2% specificity (Tab. 2). The scatter plot in Fig. 4 illustrates the strong correlation between the automated AFHC-based diagnosis and SGT.

To compare the inter-rater reliability we use Cohen’s Kappa coefficient κ [12]. The agreement between the clinicians’ GGT and SGT is 85.7% on the test data, with $\kappa=0.689$. By contrast, the AFHC-based diagnosis agrees with SGT on 89.8% of the test data, with $\kappa=0.778$. The three pairwise comparisons are summarised in Tab. 3. The agreement between all three diagnoses amounts to 81.6% of the test set, measured with a Fleiss’ Kappa of 0.734.

Tab. 3 indicates that the algorithm outperforms GGT, which mirrors the pragmatic clinical practice. The algorithm does not overly miss-predict, but may struggle to identify the precise value in cases where DDH is borderline (Fig. 4).

The correlation statistics are already high enough for our automated tool to be widely adopted for infant hip screening. It is nevertheless essential to acknowledge that the false negative cases are not to be taken lightly. In Fig. 5 we illustrate the subtle differences between individual scans from the four possible categories. Fig. 5(a) shows a true positive example, with

Table 1. Mean and standard deviation of overlap measures over the test set images. Dice Similarity Coefficient (DSC), Symmetric Boundary Dice (SBD) detecting boundary match within 3 pixels, True Positive, True Negative, False Positive, False Negative Volume Fractions (TPVF, TNVF, FPVF and FNVF) and Precision (Prec) are shown (Recall is equivalent to TPVF). Hausdorff (HD) and Average Symmetric Surface Distance (ASSD) are measured in pixels.

Metric	Ilium	Femoral Head	Labrum
DSC	0.857±0.049	0.924±0.03	0.71±0.156
SBD	0.794±0.063	0.614±0.076	0.644±0.144
TPVF	0.889±0.058	0.982±0.022	0.727±0.194
TNVF	0.996±0.002	0.979±0.011	0.998±0.001
FPVF	0.004±0.002	0.021±0.011	0.002±0.001
FNVF	0.111±0.058	0.018±0.022	0.273±0.194
Prec	0.833±0.076	0.875±0.058	0.723±0.141
HD	5.784±12.829	6.081±7.774	4.29±2.484
ASSD	0.718±0.367	1.888±1.057	1.103±0.881

Table 2. Comparison of algorithm diagnosis AFHC against SGT. True Positive Rate (TPR) is the same as Sensitivity, True Negative Rate (TNR) is the same as Specificity. FPR = 1-TNR. FNR = 1-TPR.

Correctly diagnosed with DDH (Sensitivity or TPR)	90.6%
Missed DDH (FNR)	9.4%
Correctly diagnosed no DDH (Specificity or TNR)	88.2%
Incorrectly diagnosed with DDH (FPR)	11.8%
Percentage of hips correctly diagnosed (Accuracy)	89.8%

AFHC=39.6%, an error of less than 0.04% to SGT. It is worth noting that, for 128×128 pixel images, a disagreement of one pixel represents a percentage difference of 0.78% across the whole image, and approximately 1.5% across the span of the femoral head structure. Fig. 5(b) shows a true negative example, with AFHC=62.4%, an error of less than 0.39% to SGT. Fig. 5(c) shows the most severe of the false positives. The algorithm under-segments the lowermost component of the femoral head, with AFHC=43.3%, compared to a SGT FHC of 50.4%. Fig. 5(d) shows the most severe of the false negatives. The algorithm over segments the lowermost component of the femoral head and also slightly over segments the ilium on its upper edge by including a small bump of cartilage. There is only little of the horizontal length of the ilium visible in the image; this affects the ilium edge disproportionately, and thus the algorithm calculates AFHC=60.1%, whereas the SGT FHC is 49.9% (a figure which would have caused a clinician to keep the patient under observation).

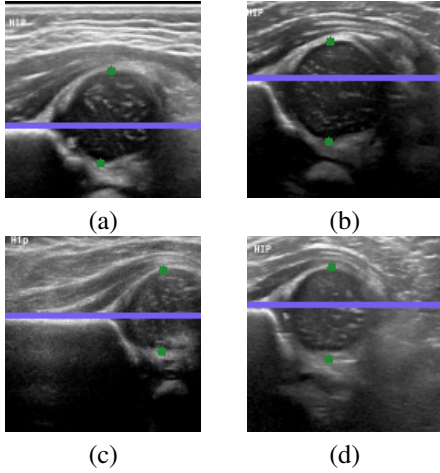


Fig. 5. Qualitative analysis of AFHC% and its corresponding diagnostic outcomes: (a) True positive. (b) True negative. (c) False positive. (d) False negative diagnosis.

Table 3. Pairwise comparison of AFHC, SGT and GGT.

Method 1 vs. Method 2	Agreement%	Cohen's Kappa
AFHC vs. SGT	89.8%	0.778
GGT vs. SGT	85.7%	0.689
AFHC vs. GGT	87.8%	0.737

5. CONCLUSION AND FUTURE WORK

To our knowledge this is the first study to estimate DDH severity by automating FHC calculation. Our algorithm was trained, validated and tested using routinely collected screening images. It outperforms the existing international state of the art for automated DDH screening using a conventional ultrasound probe. Our method makes a substantial contribution towards assistive algorithms that facilitate automating universal DDH diagnosis for every newborn.

In a screening setting, the 50% decision value could be increased thus increasing sensitivity (at the cost of lowering specificity), allowing the algorithm to conservatively filter out healthy hips, maximising the focus of specialist clinical time on DDH patients and borderline cases. If clinical trials support this, the currently prohibitive workload and cost, of universal sonographic screening becomes affordable.

Although we have not explicitly differentiated here between dysplastic and dislocated hips, this can be easily achieved through the setting of a further decision value.

When a full clinical dataset becomes available, we will compare AFHC against multiple clinicians, showing the automated diagnosis comes within inter-rater variability. We will replace the segmentation CNN with a more expressive model, thereby achieving more accurate segmentations. We expect this to improve the AFHC accuracy further.

6. ACKNOWLEDGEMENTS

The authors are grateful to the Alder Hey Children's NHS Foundation Trust who made the anonymised data available and oversaw the ethical approval for its use for research, and to Daniel Perry and Sandeep Hemmadi who provided the GGT. The study was conducted retrospectively. Each patient's parent/carer had already agreed to sharing their child's ultrasound image for this purpose. The authors have no conflicts of interest to declare.

7. REFERENCES

- [1] T Terjesen, "Residual hip dysplasia as a risk factor for osteoarthritis in 45 years follow-up of late-detected hip dislocation," *Journal of Children's Orthopaedics*, 2011.
- [2] J McCarthy *et al.*, "Developmental dysplasia of the hip," *Current Orthopaedics*, vol. 19(3), pp. 223–230, 2005.
- [3] H Dogruel *et al.*, "Clinical examination versus ultrasonography in detecting DDH," *International orthopaedics*, vol. 32(3), pp. 415–419, 2008.
- [4] C Morin *et al.*, "The infant hip: real-time us assessment of acetabular development," *Radiology*, vol. 157, no. 3, pp. 673–677, 1985.
- [5] C Gunay *et al.*, "Correlation of femoral head coverage and Graf α angle in infants being screened for developmental dysplasia of the hip," *International orthopaedics*, vol. 33(3), pp. 761–764, 2009.
- [6] S W Lee *et al.*, "Accuracy of new deep learning model-based segmentation and key-point multi-detection method for ultrasonographic DDH screening," *Diagnostics (Basel)*, 2021.
- [7] D Golan *et al.*, "Fully automating Graf's method for DDH diagnosis using deep CNNs," in *Deep Learning and Data Labeling for Medical Applications*, Cham, 2016, pp. 130–141, Springer International Publishing.
- [8] A Hareendranathan *et al.*, "Toward automatic diagnosis of hip dysplasia from 2D ultrasound," in *IEEE ISBI*, 2017, pp. 982–985.
- [9] P Cianci *et al.*, "Gestalt diagnosis for children with suspected genetic syndromes," in *Italian Journal of Pediatrics*. Springer, 2015, vol. 41(2), pp. 1–2.
- [10] O Ronneberger *et al.*, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.
- [11] UK Government, "Newborn and infant physical examination (NIPE) screening programme handbook," 2021.
- [12] L Marston, *Introductory statistics for health and nursing using SPSS*, Sage Publications, 2010.