Evaluation and Comparison of Current Fetal Ultrasound Image Segmentation Methods for Biometric Measurements: A Grand Challenge

Sylvia Rueda*, Sana Fathima, Caroline L. Knight, Mohammad Yaqub, Aris T. Papageorghiou, Bahbibi Rahmatullah, Alessandro Foi, *Senior Member, IEEE*, Matteo Maggioni, Antonietta Pepe, Jussi Tohka, Richard V. Stebbing, John E. McManigle, *Student Member, IEEE*, Anca Ciurte, Xavier Bresson, Meritxell Bach Cuadra, Changming Sun, *Member, IEEE*, Gennady V. Ponomarev, Mikhail S. Gelfand, Marat D. Kazanov, Ching-Wei Wang, *Member, IEEE*, Hsiang-Chou Chen, Chun-Wei Peng, Chu-Mei Hung, and J. Alison Noble

***Abstract*—This paper presents the evaluation results of the methods submitted to *Challenge US: Biometric Measurements from Fetal Ultrasound Images*, a segmentation challenge held at the IEEE International Symposium on Biomedical Imaging 2012. The challenge was set to compare and evaluate current fetal ultrasound image segmentation methods. It consisted of automatically segmenting fetal anatomical structures to measure standard obstetric biometric parameters, from 2D fetal ultrasound images taken on fetuses at different gestational ages (21 weeks, 28 weeks, and 33 weeks) and with varying image quality to reflect data encountered in real clinical environments. Four independent sub-challenges were proposed, according to the objects of interest measured in clinical practice: abdomen, head, femur, and whole fetus. Five teams participated in the head sub-challenge and two teams in the femur sub-challenge, including one team who tackled both. Nobody attempted the abdomen and whole fetus sub-challenges. The challenge goals were two-fold and the participants were asked to submit the segmentation results as well as the measurements derived from the segmented objects. Extensive quantitative (region-based, distance-based, and Bland–Altman measurements) and qualitative evaluation was performed to compare the results from a representative selection of current methods submitted to the challenge. Several experts (three for the head sub-challenge and two for the femur sub-challenge), with different degrees of expertise, manually delineated the objects of interest to define the ground truth used within the evaluation framework. For the head sub-challenge, several groups produced results that could be potentially used in clinical settings, with comparable performance to manual delineations. The femur sub-challenge had inferior performance to the head sub-challenge due to the fact that it is a harder segmentation problem and that the techniques presented relied more on the femur's appearance.**

***Index Terms*—Challenge, evaluation, fetal biometry, image quality, segmentation, ultrasound (US).**

Manuscript received May 29, 2013; revised July 24, 2013; accepted July 25, 2013. Date of publication August 06, 2013; date of current version March 31, 2014. *Asterisk indicates corresponding author.*

Due to space constraints, funding information for this work appears in the acknowledgement section.

S. Rueda is with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, OX3 7DQ Oxford, U.K. (e-mail: sylvia.rueda@eng.ox.ac.uk).

S. Fathima, M. Yaqub, B. Rahmatullah, R. V. Stebbing, J. E. McManigle, and J. A. Noble are with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, OX3 7DQ Oxford, U.K.

C. L. Knight and A. T. Papageorghiou are with the Nuffield Department of Obstetrics and Gynaecology, University of Oxford, OX3 7DQ Oxford, U.K.

A. Foi, M. Maggioni, A. Pepe, and J. Tohka are with the Department of Signal Processing, Tampere University of Technology, 33101 Tampere, Finland.

A. Ciurte is with the Department of Computer Science, Technical University of Cluj-Napoca, 400020 Cluj-Napoca, Romania.

M. Bach Cuadra is with the Department of Radiology, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Center for Biomedical Imaging (CIBM), and the Signal Processing Laboratory 5 (LTS5), Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland.

X. Bresson is with the Computer Science Department, City University of Hong Kong, Hong Kong.

C. Sun is with the CSIRO Computational Informatics, North Ryde, NSW 1670, Australia.

G. V. Ponomarev, M. S. Gelfand, and M. D. Kazanov are with the Research and Training Center on Bioinformatics, Institute for Information Transmission Problems, 127994 Moscow, Russia.

C.-W. Wang, H.-C. Chen, C.-W. Peng, and C.-M. Hung are with the Graduate Institute of Biomedical Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2013.2276943

I. Introduction

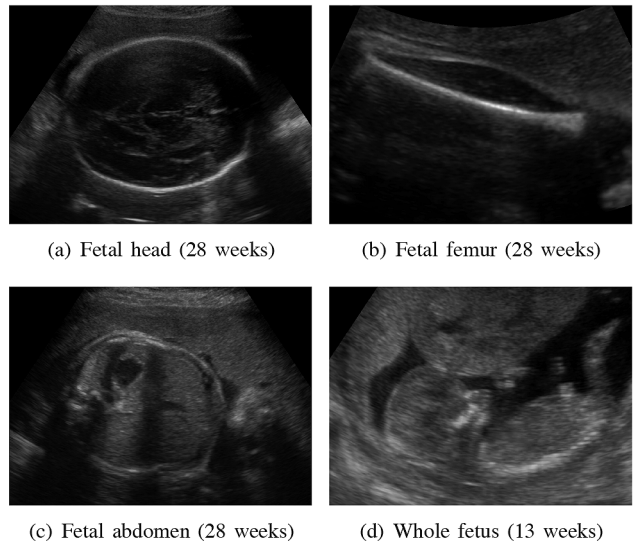
ULTRASOUND (US) imaging is the modality of choice in many clinical applications due to its non-invasive nature, reduced cost, and real-time acquisition, compared to other imaging modalities, such as computed tomography (CT) or magnetic resonance imaging (MRI). However, US images are patient-specific, operator-dependent, and machine specific, which makes image appearance tightly linked to patient characteristics, the expertise of the clinician acquiring the images, and the machine used. Besides, due to the properties of image formation intrinsic to US images, they can be affected by signal dropouts, artefacts, missing boundaries, attenuation, shadows, and speckle, making US one of the most challenging modalities to work with. Depending on the orientation of the transducer, the image obtained might not have the expected anatomical significance and can be distorted or incomplete. Protocols are defined to acquire the best possible images while retaining the characteristics of the object of interest (e.g., shape and anatomy).

2D fetal US biometrics have been extensively used to establish (or confirm) the gestational age of the fetus, estimate

its size and weight, and identify growth patterns and abnormalities [1]. Typically, fetal size is estimated by using 2D US measurements of head, abdomen, and femur, at around 20 weeks gestational age [2]. These measurements, and any at later gestations, are then compared with population-based growth charts to identify normal or abnormal growth. In an attempt to reduce intra- and inter-observer variability, and create more accurate and reproducible measurements [3], [4], automatic methods for fetal biometric measurements have been investigated recently. Furthermore, automated fetal biometry has been shown to improve the work flow efficiency by reducing the examination time and the number of steps necessary for standard fetal measurements [5]. This would also benefit less experienced users.

It is worth noting that automated analysis of US images is hard, and methods developed for MRI and CT do not necessarily work on US images. Furthermore, general methods for US image segmentation do not exist, and the segmentation strategies are application dependent [6]. The automatic segmentation methods previously developed in the fetal imaging field focused on using segmentation as an intermediate processing step for estimating standard biometric measurements. Most of the methods attempted to segment the fetal femur [7]–[10], the fetal head [11]–[16], or both [17]–[19]. The methods were based on morphological operators, active contour models, Hough transform, deformable models, or machine learning approaches. Low level features and textures were frequently used to find the femur and the skull, because these have a brighter response [Fig. 1(a) and (b)]. However, the task of segmenting the abdomen is more challenging and only few works have attempted it up-to-date [14], [20], [21]. General methods retrieving all standard fetal biometric measurements used in antenatal clinical practice are limited [22], [23], [4], [24]. Carneiro *et al.* [22] used a discriminative constrained probabilistic boosting tree classifier to segment structures of interest and to reproduce standard biometric measurements for all three objects of interest (head, abdomen, and femur) in fetal US images. They developed and patented a commercial system, called Auto OB [4], which is integrated into Siemens software and that can detect, apart from head, abdomen, and femur biometric measurements, the humerus length (HL) and the crown-rump length (CRL). This is the only system for fetal biometry that has been translated into clinical practice.

Among the different objects of interest, the simplest segmentation and detection appears to be the head [Fig. 1(a)], because it presents clear boundaries and texture similarities among individuals The fetal femur [Fig. 1(b)] can lack internal texture, which can make its accurate delineation difficult, but most of the time strong edges are present in most of their contour except in the extremities. The abdomen [Fig. 1(c)] and the whole fetus [Fig. 1(d)] segmentations are the hardest because they lack clear boundaries and have inconsistencies in the internal structures among individuals. Furthermore, the healthy fetal body changes its shape across gestation, as a result of growth, and the different organs that surround the object of interest create high pose and shape variability for the same structure.



(a) Fetal head (28 weeks) (b) Fetal femur (28 weeks)

(c) Fetal abdomen (28 weeks) (d) Whole fetus (13 weeks)

Fig. 1. Ultrasound images of (a) the fetal head, (b) the fetal femur, (c) the fetal abdomen, and (d) the whole fetus.

This paper presents the evaluation and comparison of the representative selection of current methods presented during *Challenge US: Biometric Measurements from Fetal Ultrasound Images*[1], a segmentation challenge held in conjunction and with the support of the IEEE International Symposium on Biomedical Imaging (ISBI) 2012. The challenge consisted of four independent sub-challenges according to the objects of interest measured in clinical practice on 2D fetal ultrasound images: abdomen, head, femur, and whole fetus (Fig. 1). The images were selected at three different gestational ages (21 weeks, 28 weeks, and 33 weeks) and with varying image quality to represent real clinical environments. The gestational ages were selected from 20 weeks onwards, as this is representative of a real clinical setting for this particular application. Several experts, with different degrees of expertise, manually delineated the objects of interest to define the ground truth, which was used within the segmentation framework. Extensive quantitative and qualitative evaluation was performed to assess the performance of the methods with respect to manual delineations.

Apart from the segmentation results, participants were asked to estimate biometric measurements derived from the segmented objects, which are the values used clinically for fetal growth assessment. The evaluation of the segmentation results and derived measurements were performed separately, since a segmentation result can be poor and still lead to good measurements. One key aspect missing in most US strategies is the ability to incorporate image quality within the comparison, to understand which methods are more susceptible to changes in appearance. We have deliberately included analysis on data of different degrees of difficulty to better understand degradation of methods with quality.

Five teams participated in the head sub-challenge and two teams in the femur sub-challenge, including one team who tackled both. Nobody attempted the abdomen and the whole

[1]http://www.ibme.ox.ac.uk/challengeus2012

fetus sub-challenges. This is to our knowledge the first segmentation challenge undertaken in the fetal US imaging field, and thus provides both a reference publication from which to gauge how well a representative selection of current methods work today and may encourage others to work in this area.

In Section II, we introduce the challenge aims, the description of image data sets used within the challenge, and the description of the fetal biometric measurements for the structures of interest. Section III presents the evaluation metrics used to compare the segmentation results and derived measurements. Section IV introduces the ground truth and its reproducibility study. Section V summarizes the methodologies presented to the challenge. Quantitative and qualitative results are described in Section VI. A discussion and conclusions are given in Sections VII and VIII, respectively.

II. Challenge US: Biometric Measurements From Fetal Ultrasound Images

A. Organization

The challenge was set up to automatically segment anatomical structures to measure standard obstetric biometric parameters, from 2D fetal ultrasound images, taken on fetuses at different gestational ages (21 weeks, 28 weeks, and 33 weeks). The segmentation challenge was formed by 4 sub-challenges, named fetal head, fetal abdomen, fetal femur, and whole fetus. The participation was open to those wanting to attempt one or several of these sub-challenges, presenting different degrees of difficulty. General solutions applicable to all four sub-challenges had more value if the performance was good. Only methods based on automatic or semi-automatic segmentation techniques were considered. The challenge was open to teams from academia and industry. Published methods were allowed to be submitted. The results from each team were automatically compared to the ground truth, obtained from expert manual segmentations and measurements. The challenge goals were two-fold, since segmented objects and derived clinical measurements were both considered to assess the quality of the methods. Two months were given to develop the methods and submit the results.

B. Description of Image Data Sets

All the images from this study were acquired by trained clinicians using the same mid-range ultrasound machine Philips HD9 and following the protocols defined by the INTERGROWTH-21st study [25]. Most of the images were acquired with a 7-3 MHz transducer. In case of later gestations or mothers having a high body mass index, the 5-2 MHz transducer was preferred. The images were in DICOM format, anonymised, and automatically cropped (to remove the header) to a size of 756×546 pixels before distribution. Spatial resolution (in millimeters) varied among the images.

Fetal head, abdomen, and femur sub-challenges had a total of 90 images each in anonymised DICOM format and the whole fetus sub-challenge a total of 14 images, as these were not routinely acquired on site. Three different gestational ages were considered at 21, 28, and 33 weeks with a total of 30 images per gestational age for each of the structures considered. The gestational ages to include in this challenge have been carefully selected after clinical advice, providing a good representation of the challenges encountered across gestation. Furthermore, for each gestational age, three groups of different qualities were obtained. These were graded as low, medium, and high quality and were selected as objectively as possible to create real image data sets as used in clinical practice. The reader is referred to Appendix A for details on the image scoring criteria used within this framework.

C. Participation in the Challenge

A total of six teams submitted results to the challenge. Five teams participated in the fetal head sub-challenge:

Foi *et al.* [26], Head contour extraction from the fetal ultrasound images by difference of Gaussians revolved along elliptical paths. (Finland).

Ciurte *et al.* [27], A semi-supervised patch-based approach for segmentation of fetal ultrasound imaging. (Switzerland).

Stebbing and McManigle [28], A boundary fragment model for head segmentation in fetal ultrasound. (U.K.).

Sun [29], Automatic fetal head measurements from ultrasound images using circular shortest paths. (Australia).

Ponomarev *et al.* [30], A multilevel thresholding combined with edge detection and shape-based recognition for segmentation of fetal ultrasound images. (Russia).

Two teams participated in the femur sub-challenge:

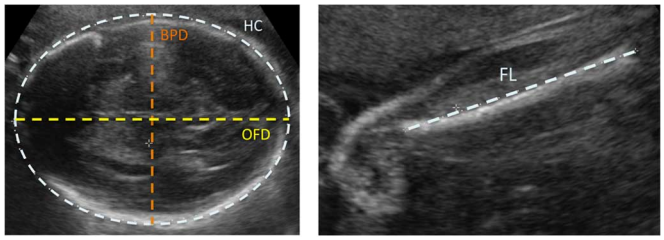
Ponomarev *et al.* [30], A multilevel thresholding combined with edge detection and shape-based recognition for segmentation of fetal ultrasound images. (Russia).

Wang *et al.* [31], Automatic femur segmentation and length measurement from fetal ultrasound images. (Taiwan).

Only the method by Ponomarev *et al.* [30] attempted to solve both sub-challenges simultaneously. No attempts were made on abdomen and whole fetus segmentations. This could be due to the fact that these two sub-challenges were harder because the images tend to have fuzzy boundaries and present inconsistencies in the internal structures among individuals. Another possible explanation would be the limited amount of time the teams had to develop a new method. In the rest of the paper, we will only focus on the head and femur sub-challenges.

D. Standard Fetal Biometry

Three standard fetal biometric measurements of the head were considered: Biparietal Diameter (BPD), Occipito-Frontal Diameter (OFD), and Head Circumference (HC), as shown in Fig. 2(a). Several ways of measuring BPD and OFD exist (e.g., outer-to-outer, inner-to-outer). In this paper, BPD and OFD are defined as in the INTERGROWTH-21st study [25]. These measures are shown in Fig. 2(a). The HC parameter is derived from BPD and OFD parameters as $\mathrm{HC} = \pi(\mathrm{BPD} + \mathrm{OFD})/2$. Another standard measure for fetal biometry consists of measuring the femur length (FL). The FL is measured from the outer



(a) (b)

Fig. 2. (a) Fetal Head Biometric Measurements: Head Circumference (HC), Biparietal Diameter (BPD), and Occipito-Frontal Diameter (OFD). (b) Fetal Femur Biometric Measurement: Femur Length (FL).

TABLE I
RESULTS REQUIRED FOR EACH SUB-CHALLENGE

Sub-Challenge	Segmented Object	Ellipses from Measurements	Biometric Measurements
Head		√	BPD, OFD, HC
Femur	√		FL

edges of the bone, without taking into account the trochanter of the femur, as shown in Fig. 2(b).

E. Submission of Results

The results submitted depended on the sub-challenge attempted, as summarised in Table I. For the fetal head, due to the huge difficulty in manually delineating the actual objects in a variety of ultrasound images, the binary image resulting from the ellipse fitted object was used as the result. The value for the binary image pixels on the contour and inside of the ellipses needed to be equal to 1 (foreground) and the rest equal to 0 (background).

For the fetal femur, the whole segmented structure needed to be obtained as part of the segmentation challenge. Recent clinical evidence [32], [33] has shown that other femoral characteristics, apart from the femur length, are important to assess fetal bone growth and development. Automatic and accurate tools for whole femur bone segmentation, although limited, have shown great potential [34], [35] and are able to perform more complex measurements for a better fetal bone development assessment. This is the clinical motivation for incorporating whole femur bone segmentation into this challenge.

From the segmented objects, the biometric measurements could be derived and needed to be presented as part of the results, with the binary images. The measurements needed to be reported in millimeters, using the DICOM information providing the resolution of each image.

III. EVALUATION METRICS

The evaluation metrics chosen attempt to assess the quality of the segmentation as well as the measurements. Three different criteria were considered. First, region-based metrics were selected to assess the precision, specificity, sensitivity, and Dice similarity. Then, distance-based metrics were used to quantify the local variability existing between the proposed methods and manual delineations. Finally, Bland–Altman plots were used to compare against clinical measurements, to show the agreement between the proposed methods and the experts. These metrics are defined in the following.

A. Region-Based Metrics

Region-based evaluation metrics, as defined in [36], were selected as a way of assessing precision and accuracy of different segmentation methods. Due to the difficulty of establishing true segmentations, segmentation results were compared to manual delineations of the structures, performed by several operators twice on each image. The results per image were averaged to obtain the overall performance for a particular expert and for all experts. In the following, let $O_{\mathrm{SR}}^{\mathrm{M}}$ denote the segmentation results for a method M and O_{GT} the ground truth delineated by the experts. All region-based metrics are given as percentages.

1) Precision: The precision P assesses the reproducibility of each segmentation method. P characterizes the common amount of tissue in both $O_{\mathrm{SR}}^{\mathrm{M}}$ and O_{GT} as a fraction of the total amount of tissue in the union of $O_{\mathrm{SR}}^{\mathrm{M}}$ and O_{GT} as

$$P = \frac{\left|O_{\mathrm{SR}}^{\mathrm{M}} \cap O_{\mathrm{GT}}\right|}{\left|O_{\mathrm{SR}}^{\mathrm{M}} \cup O_{\mathrm{GT}}\right|}. \tag{1}$$

2) Accuracy: True positive (TP) and true negative (TN) measures are calculated to assess the accuracy of each method [36]. TP is the fraction of the total amount of tissue in the true delineation that was covered by the method and represents the *delineation sensitivity*. It is defined as

$$\mathrm{TP} = \frac{\left|O_{\mathrm{SR}}^{\mathrm{M}} \cap O_{\mathrm{GT}}\right|}{\left|O_{\mathrm{GT}}\right|}. \tag{2}$$

TN is the fraction of the total amount of tissue in the reference region U that does not belong to the object and was excluded from the method. It represents the *delineation specificity* and is defined as

$$\mathrm{TN} = \frac{\left|\left(O_{\mathrm{SR}}^{\mathrm{M}} \cup O_{\mathrm{GT}}\right)^{c}\right|}{\left|(O_{\mathrm{GT}})^{c}\right|} \tag{3}$$

where $(\cdot)^c$ denotes the absolute complement of a set for a fixed reference region U. The greater the TN values, the better the delineation accuracy of a method.

3) Dice Similarity: Dice similarity D gives an indication of the mutual overlap between $O_{\mathrm{SR}}^{\mathrm{M}}$ and O_{GT}. D is defined as

$$D = \frac{2\left|O_{\mathrm{GT}} \cap O_{\mathrm{SR}}^{\mathrm{M}}\right|}{\left|O_{\mathrm{GT}}\right| + \left|O_{\mathrm{SR}}^{\mathrm{M}}\right|}. \tag{4}$$

B. Distance-Based Metrics

Along with area overlap measures defined previously, distance-based metrics, as described in [37], are incorporated into the evaluation to provide different ways of assessing the errors of the different segmentation methods. These measures are given in millimeters.

1) Maximum Symmetric Contour Distance: Let $\mathcal{C}(O_{\mathrm{GT}})$ and $\mathcal{C}(O_{\mathrm{SR}}^{\mathrm{M}})$ be the contours of O_{GT} and $O_{\mathrm{SR}}^{\mathrm{M}}$, respectively. $c_{O_{\mathrm{GT}}}$

denotes a contour element of $\mathcal{C}(O_{\mathrm{GT}})$ and $c_{O_{\mathrm{SR}}^{\mathrm{M}}}$ a contour element of $\mathcal{C}(O_{\mathrm{SR}}^{\mathrm{M}})$. The shortest distance of a pixel p to $\mathcal{C}(O_{\mathrm{GT}})$ is defined as

$$d_E(p, \mathcal{C}(O_{\mathrm{GT}})) = \min_{c_{O_{\mathrm{GT}}} \in \mathcal{C}(O_{\mathrm{GT}})} \|p - c_{O_{\mathrm{GT}}}\| \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean distance. The Maximum Symmetric Contour Distance (MSD), also known as Hausdorff distance [38], can then be expressed as

$$\mathrm{MSD}\left(O_{\mathrm{GT}}, O_{\mathrm{SR}}^{\mathrm{M}}\right) = \max\left(\max_{c_{O_{\mathrm{GT}}} \in \mathcal{C}(O_{\mathrm{GT}})} d_E\left(c_{O_{\mathrm{GT}}}, \mathcal{C}\left(O_{\mathrm{SR}}^{\mathrm{M}}\right)\right) \max_{c_{O_{\mathrm{SR}}^{\mathrm{M}}} \in \mathcal{C}(O_{\mathrm{SR}}^{\mathrm{M}})} d_E\left(c_{O_{\mathrm{SR}}^{\mathrm{M}}}, \mathcal{C}(O_{\mathrm{GT}})\right)\right). \tag{6}$$

This measure is sensitive to outliers and returns the maximum error, which represents the worst case scenario.

2) Average Symmetric Contour Distance: The Average Symmetric Contour Distance (ASD) corresponds to the average of all distances between O_{GT} and $O_{\mathrm{SR}}^{\mathrm{M}}$ defined as

$$\mathrm{ASD}\left(O_{\mathrm{GT}}, O_{\mathrm{SR}}^{\mathrm{M}}\right) = \frac{1}{|\mathcal{C}(O_{\mathrm{GT}})| + \left|\mathcal{C}\left(O_{\mathrm{SR}}^{\mathrm{M}}\right)\right|} \times \left(\sum_{c_{O_{\mathrm{GT}}} \in \mathcal{C}(O_{\mathrm{GT}})} d_E\left(c_{O_{\mathrm{GT}}}, \mathcal{C}\left(O_{\mathrm{SR}}^{\mathrm{M}}\right)\right) + \sum_{c_{O_{\mathrm{SR}}^{\mathrm{M}}} \in \mathcal{C}(O_{\mathrm{SR}}^{\mathrm{M}})} d_E\left(c_{O_{\mathrm{SR}}^{\mathrm{M}}}, \mathcal{C}(O_{\mathrm{GT}})\right)\right) \tag{7}$$

where $|\cdot|$ denotes the length of the contour. A perfect segmentation would return a value of 0 mm.

3) Root Mean Square Symmetric Contour Distance: The Root Mean Square Symmetric Contour Distance (RMSD) is defined in (8), as shown at the bottom of the page. The RMSD is similar to the ASD but large distance differences between contours will return a greater value, penalizing large deviations from the ground truth.

C. Bland–Altman Plots

Bland–Altman plots [39], [40] assess the agreement between two sets of measurements. In this study, Bland–Altman plots are used to compare the measurements derived from the segmentation results to the clinical measurements performed by the different experts. This technique can also be used to obtain the inter- and intra-observer variability measurements.

D. Efficiency

Average segmentation times, software, and hardware used by each method are reported in the paper but none of the methods had been implemented for efficiency so such times are not a guide to practical deployment.

E. Failures

The failures of a method are reported individually on each image when no overlap exists between the segmentation result and the ground truth delineated by the experts. Failures are excluded from the segmentation evaluation and reported separately.

IV. Ground Truth and Its Reproducibility

A. Fetal Head Sub-Challenge

A total of three experts, with different degrees of expertise, participated in defining the fetal head sub-challenge ground truth, by fitting an ellipse to the object of interest twice on each image, as well as performing the corresponding standard clinical measurements (HC, BPD, OFD). The experts for the head sub-challenge had the following level of expertise.

- **Expert 1**: Clinician (fetal medicine specialist) with 10 year postgraduate experience in fetal US scans.
- **Expert 2**: Clinician (obstetrician) with two years experience in fetal US scans.
- **Expert 3**: Engineer with 1 year of experience.

The intra- and inter-observer variability was calculated independently for each expert using the metrics defined in Section III. The average intra-expert variability results (resulting from comparing manual delineations) over all images are presented in Table II. The intra-expert variability is similar for all three experts. Although there were minor differences reflecting the levels of experience, these were not statistically significant. Expert 3, who was the less experienced, obtained slightly inferior results than the other two experts, but still very close.

The average inter-expert variability results over all images are presented in Table III comparing the manual delineations from different experts two by two. The results are very similar between all combinations of experts.

The intra- and inter-expert variability of the fetal biometric measurements can be assessed using Bland–Altman plots, as reported in Tables IV and V, respectively. The mean values

$$\mathrm{RMSD}\left(O_{\mathrm{GT}}, O_{\mathrm{SR}}^{\mathrm{M}}\right) = \sqrt{\frac{1}{|\mathcal{C}(O_{\mathrm{GT}})| + \left|\mathcal{C}\left(O_{\mathrm{SR}}^{\mathrm{M}}\right)\right|}} \times \sqrt{\sum_{c_{O_{\mathrm{GT}}} \in \mathcal{C}(O_{\mathrm{GT}})} d_E^2\left(c_{O_{\mathrm{GT}}}, \mathcal{C}(O_{\mathrm{SR}}^{\mathrm{M}})\right) + \sum_{c_{O_{\mathrm{SR}}^{\mathrm{M}}} \in \mathcal{C}\left(O_{\mathrm{SR}}^{\mathrm{M}}\right)} d_E^2\left(c_{O_{\mathrm{SR}}^{\mathrm{M}}}, \mathcal{C}(O_{\mathrm{GT}})\right)} \tag{8}$$

TABLE II
INTRA-OBSERVER VARIABILITY OF MANUAL DELINEATIONS: FETAL HEAD

Metric	Expert 1	Expert 2	Expert 3
Precision (%)	96.54 ± 1.38	96.64 ± 1.46	96.11 ± 1.79
Sensitivity (%)	97.81 ± 1.38	97.93 ± 1.58	98.90 ± 1.46
Specificity (%)	99.24 ± 0.82	99.26 ± 0.69	98.37 ± 1.25
Dice (%)	98.24 ± 0.71	98.28 ± 0.76	98.01 ± 0.94
MSD (mm)	1.72 ± 0.81	1.74 ± 1.09	1.85 ± 1.10
ASD (mm)	0.69 ± 0.32	0.68 ± 0.35	0.79 ± 0.44
RMSD (mm)	0.85 ± 0.39	0.83 ± 0.47	0.95 ± 0.54

TABLE III
INTER-OBSERVER VARIABILITY OF MANUAL DELINEATIONS: FETAL HEAD

(E1: Expert 1 – E2: Expert 2 – E3: Expert3)

Metric	E1 vs E2	E2 vs E3	E1 vs E3
Precision (%)	95.84 ± 1.39	95.45 ± 1.46	95.78 ± 1.48
Sensitivity (%)	97.46 ± 1.30	98.34 ± 1.31	98.08 ± 1.17
Specificity (%)	99.00 ± 1.01	98.28 ± 0.99	98.59 ± 1.11
Dice (%)	97.87 ± 0.73	97.66 ± 0.77	97.83 ± 0.78
MSD (mm)	2.11 ± 1.12	2.24 ± 1.19	2.09 ± 0.99
ASD (mm)	0.86 ± 0.39	0.93 ± 0.42	0.86 ± 0.40
RMSD (mm)	1.04 ± 0.49	1.13 ± 0.54	1.05 ± 0.49

TABLE IV
INTRA-OBSERVER VARIABILITY OF CLINICAL MEASUREMENTS: FETAL HEAD

Measure	Expert 1	Expert 2	Expert 3
BPD (mm)	0.31 ± 1.57	-0.04 ± 0.54	0.13 ± 0.79
OFD (mm)	0.64 ± 1.99	0.82 ± 1.98	-0.67 ± 1.98
HC (mm)	1.1 ± 3.14	1.23 ± 3.31	-2.53 ± 4.19

TABLE V
INTER-OBSERVER VARIABILITY OF CLINICAL MEASUREMENTS: FETAL HEAD

(E1: Expert 1 – E2: Expert 2 – E3: Expert 3)

Measure	E1 vs E2	E2 vs E3	E1 vs E3
BPD (mm)	0.39 ± 1.66	-0.47 ± 0.89	-0.08 ± 1.84
OFD (mm)	-1.55 ± 2.36	1.09 ± 2.75	-0.45 ± 2.24
HC (mm)	0.68 ± 4.15	0.65 ± 3.76	1.33 ± 4.07

in Table IV correspond to the bias between both measurements for each expert. The standard deviations represent the random error existing between measurements (reproducibility). Standard deviations in Table V represent the reproducibility of the measurements between experts. Both the intra- (Table IV) and inter-expert (Table V) variability have a lower standard deviation than previously reported values [41], [42], indicating a higher reproducibility. This is due to the fact that this study was performed on a different clinical database to the ones used in [41] and [42] and that the experts had different levels of expertise. In the remaining of the paper, the reproducibility of the biometric measurements submitted to the head sub-challenge will be compared to those reported in Tables IV and V.

B. Fetal Femur Sub-Challenge

For the femur sub-challenge, two experts performed manual delineation of the fetal femur and measured the FL twice on each image. Delineation of the whole femur is not done in routine clinical practice, therefore only two experts were considered in this case to account for manual tracing variability, whereas more clinicians are experienced in biometric measurements. The experts had the following level of expertise.

- **Expert 1**: Engineer with more than three years of experience in fetal femur segmentation.
- **Expert 2**: Clinician (obstetrician) with two years experience in fetal US scans.

TABLE VI
INTRA- AND INTER-OBSERVER VARIABILITY OF MANUAL DELINEATIONS: FEMUR

(E1: Expert 1 – E2: Expert 2)

	Intra-Observer Variability		Inter-Observer Variability
Metric	**E1**	**E2**	**E1 vs E2**
Precision (%)	79.55 ± 5.25	77.52 ± 5.57	73.52 ± 5.78
Sensitivity (%)	88.48 ± 6.45	86.90 ± 6.59	78.38 ± 7.35
Specificity (%)	99.69 ± 0.21	99.71 ± 0.18	99.82 ± 0.14
Dice (%)	88.51 ± 3.32	87.22 ± 3.56	84.55 ± 3.92
MSD (mm)	1.53 ± 0.62	1.57 ± 0.95	1.93 ± 0.79
ASD (mm)	0.32 ± 0.10	0.31 ± 0.10	0.41 ± 0.10
RMSD (mm)	0.43 ± 0.14	0.42 ± 0.16	0.55 ± 0.16

TABLE VII
INTRA- AND INTER-OBSERVER VARIABILITY OF CLINICAL MEASUREMENTS: FEMUR LENGTH

(E1: Expert 1 – E2: Expert 2)

	Intra-Observer Variability		Inter-Observer
Measure	**E1**	**E2**	**E1 vs E2**
FL (mm)	-0.08 ± 1.04	-0.2 ± 1.12	-1.27 ± 1.47

Intra- and inter-expert variability are presented in Table VI. Both experts present similar results for all the metrics used. The results are inferior to those presented for the head, because the accurate delineation of the structures is more challenging and subjected to higher variability due to the fuzzy boundaries and presence of artefacts.

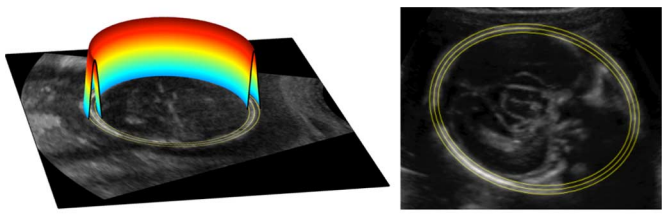
The intra- and inter-expert variability of the fetal biometric measurements can be assessed using Bland–Altman plots, as reported in Table VII. Similarly to the fetal head sub-challenge, the intra- and inter-expert variability (Table VII) show a higher reproducibility than those reported in [41], [42]. The FL measurements submitted to the femur sub-challenge will be assessed based on Table VII.

V. METHODS

This section summarizes the methods that were submitted to the different sub-challenges. For more details, we refer the reader to the individual papers.

A. Fetal Head Sub-Challenge

Five very different methods were submitted to the fetal head sub-challenge. The Foi *et al.* method [26] used signal processing operations combined with an optimization framework. The methods of Ciurte *et al.* [27] and Sun [29] used graph-based approaches. Stebbing and McManigle [28] used a machine learning approach based on a boundary fragment model resulting from a training step. The Ponomarev *et al.* method [30] defined multiple thresholds combined with edge detection and shape-based recognition and then fitted an ellipse



(a) (b)

Fig. 3. (a) Surface modelling the fetal skull by revolving a difference of Gaussians along the elliptical path. Negative parts of the surface are not visible, hidden by the US image. (b) Example on a 21 week fetus using the proposed approach. The central ellipse is the fitted ellipse. The outer ellipse is used for OFD and BPD measurements.

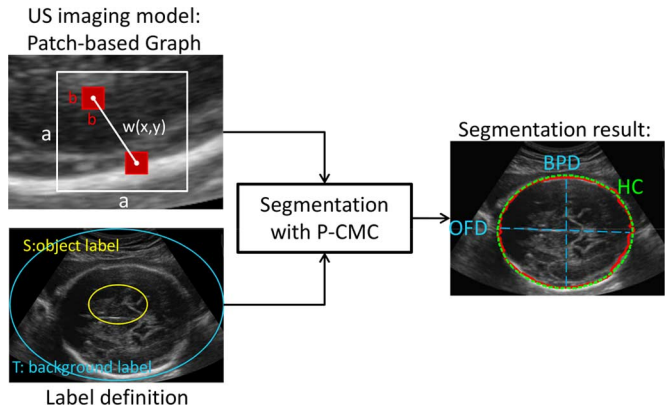
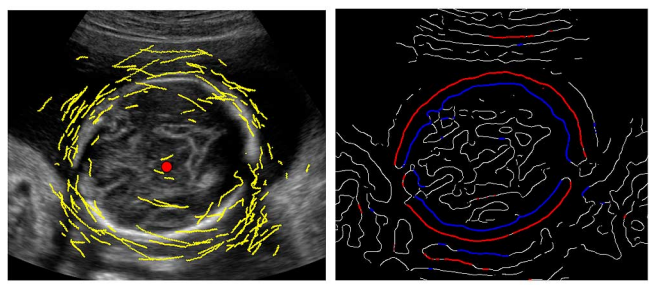


Fig. 4. Block diagram of the Patch-based Continuous Min-Cut (P-CMC) segmentation for fetal ultrasound images. Fetus of 28 weeks of gestational age. x and y correspond to two different pixels in the image (nodes of the graph). a is the size of the searching window and b represents the patch size. w(x,y) is the similarity measure between pixels x and y.

to the resulting binary image. A summary of each method is presented in the following.

1) Head Contour Extraction by Difference of Gaussians Revolved Along Elliptical Paths: Foi *et al.* [26] proposed a fully automatic method based on fitting an ellipse to each US image by modelling the fetal head contour. This was achieved by minimizing a cost function with respect to the parameters of the ellipse, by using a global multi-scale multi-start Nelder–Mead algorithm [43]. The images are first preprocessed to fill in the black background outside the scanned area by extrapolating the image inside the scanned area using a constrained iterative low-pass filter in the discrete cosine transform (DCT) domain. Then, image contrast and intensity are regularized by leveraging DCT-domain smoothing in order to provide smoothly varying local normalization of intensities. For a given ellipse, the surface that models the skull of the fetus is obtained by revolving a difference of Gaussians along the elliptical path, as shown in Fig. 3(a). The cost function can then be defined as the product of the image and the surface integrated over the image domain. The cost function is minimized globally using a multiscale multistart Nelder–Mead algorithm. The convergence of the optimization algorithm is accelerated by using a coarse-to-fine multi-scale approach, starting the process at a lower resolution and using the result to initialise higher resolutions. The final biometric measurements are derived from the major and minor axes after obtaining outer-to-outer measures of the skull. Fig. 3(b) shows the fitted ellipse and the inner and outer ellipses after incorporating the skull thickness. The method did not require any tuning of parameters.

2) Semi-Supervised Patch-Based Approach: Ciurte *et al.* [27] proposed a semi-supervised patch-based segmentation approach based on a previous work [44]. Each US image is represented by a graph of image patches (Fig. 4). A continuous min-cut partition [45] of the graph and a fast minimization scheme solve the segmentation problem. The method is semi-supervised, and therefore initial labels have to be defined on each image, to act as soft priors. In general, the labels are defined by doing a few clicks on the image, resulting in an initial polygonal shape. The automatization of the initialization was performed by setting two concentric elliptic labels at the middle of the image, as shown in Fig. 4. This assumes that the head is always at the center of the image, which is not true

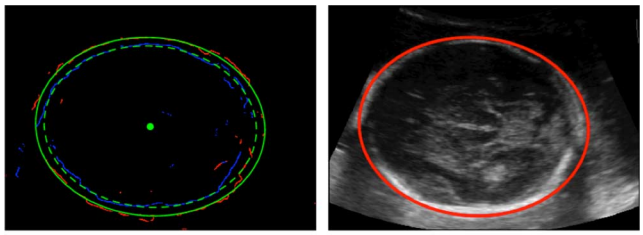


(a) Original image. (b) Final edge classification.

Fig. 5. (a) Original image with edge fragments overlaid (yellow segments). (b) Edge map derived from feature asymmetry with final edge classification overlaid (blue: inner boundary; red: outer boundary).

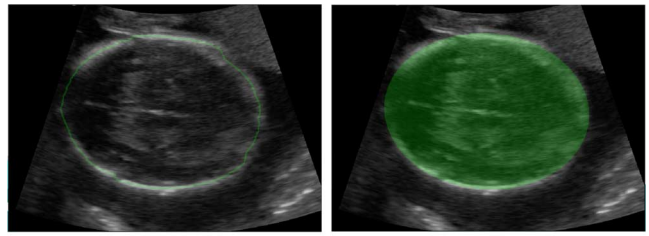
in all cases. Otherwise, manual initialization was necessary. This was the case for around half of the images in the data set. The segmentation returns a binary object with irregular contour (red contour in Fig. 4), which is used in a second step to determine its corresponding elliptical binary object. For this purpose, the axis of elongation [46] of the resulting object (or axis of least second-order moment) is computed. The elongation axis corresponds to the OFD measurement, and the BPD can be computed perpendicularly to it, for the same center of mass. An example of the resulting ellipse is given in Fig. 4 (green contour). The parameter setting was constant for all tests ($a = 5$, $b = 3$, scaling factor $\sigma = 0.004$, and regularization term $\beta = 0.001$).

3) A Boundary Fragment Model for Random Forest Edge Classification: Stebbing and McManigle [28] proposed an automatic method, based on a boundary fragment model, constructed using a machine learning approach, extending previous work [47]. The method relies only on edge information, derived from feature asymmetry [48]. From the edges, the position and orientation of edge pixels can be retrieved. A boundary fragment model [Fig. 5(a)] is then used to determine the centroid and scale of the skull by using a boosted classifier [49], which allows to identify the optimal centroid and scale of the fetal skull by using a mean-shift method. The same boundary fragment



(a) (b)

Fig. 6. Ellipse fitting step. (a) A dual ellipse fitted to inner (blue) and outer (red) contours. (b) Final result used for biometric measurements (red: outer contour).



(a) (b)

Fig. 7. (a) Closed contour (green) resulting from the CSP algorithm. (b) Final fitted ellipse overlaid to the original image.

(a) (b) (c)

Fig. 8. (a) Original image. (b) Preliminary segmented objects. (c) Inscribed head ellipse.

model is then used in a Random Forest framework to differentiate between inner, outer edges, and background [Fig. 5(b)]. An iterative dual ellipse fitting step is used to find the best inner and outer skull ellipses (Fig. 6) to derive the biometric measurements. The training samples were obtained from a set of images different from the challenge data set. Half of the training set was used to build the boundary fragment model and the other half was used to train the detection and delineation classifiers. The training data was split in half randomly, only once. The parameters used within the random forest framework were set empirically. Those needed to create the boundary fragment model were selected in line with [49]. Most of these parameters have little impact on the final performance and can be set within a wide range.

4) Circular Shortest Paths: Sun [29] proposed an automatic method based on a graph-based approach called circular shortest paths (CSP), developed in previous work [50]. The method is divided into three main steps: circular shortest path extraction, robust ellipse fitting, and finding the outer edge of the skull. The CSP algorithm ensures a closed boundary by forcing the starting and ending points of a shortest path to meet. The summation of pixel values along the object boundary is maximized to obtain the optimal path. The CSP algorithm is run up to three times. The third time will only be in the rare cases where BPD and OFD values are greater than a threshold. For each iteration, the image is converted to polar coordinates, a CSP is found, and an ellipse is fitted. The robust ellipse fitting relies only on the 50% brightest pixels on the circular shortest path, which are most likely to belong to the skull. When the CSP is run for the third time, the ellipse center is selected and the side of the ellipse which best fits the data is used to constraint the location and scale of the ellipse. Within the new constrained region, the CSP is run again and the new ellipse is found. The outer edge of the skull is then retrieved by calculating the image gradient in the radial direction in the neighborhood of the fitted ellipse boundary, pointing towards the outside of the skull, and finding an edge. The resulting edge offset can then be added to the fitted ellipse to find the outer edge of the skull to derive the biometric measurements. An example is given in Fig. 7. The parameter settings involved defining an image center, which was initially used for CSP finding; using the top 50% brightest pixels along the resulting CSP to fit the ellipse; and fixing the upper limits of BPD and OFD values to 90 and 105, respectively, for the third CSP pass. The parameter setting was constant for all tests.

5) A Multilevel Thresholding Combined With Edge Detection and Shape-Based Recognition: Ponomarev *et al.* [30] used a multilevel thresholding approach to segment the fetal skull combined with edge detection and shape-based recognition. This approach makes use of the difference in intensities between the bone and the image background, and assumes that hard tissue (bone) appears brighter than the surrounding objects in the US images. The methodology is based on multiple intensity level thresholds. For each binary image obtained, the connected components are retrieved and a measure of thinness and elongation is calculated. The candidate objects are found after applying empirically chosen thresholds. A size constraint was also applied to remove small objects. The objects resulting from the multithresholding were grouped into a cluster from which mean edge contrast was calculated to estimate the best object intensity representation. The result for each cluster was transformed into a binary image, as shown in Fig. 8. The binary image contains spurious objects due to other structures appearing in the images. Ellipses are then fitted considering all possible combinations using a scoring function, created to study the contrast around the ellipse contour, which should normally correspond to the skull. All the thresholds used within this approach were empirically chosen and fixed for all experiments. This method was also applied to the femur sub-challenge. The adaptations to this other object are defined in Section V-B.

B. Fetal Femur Sub-Challenge

Two teams participated in this sub-challenge. Both methods relied on appearance and edge information extracted directly from intensity values.

1) A Multilevel Thresholding Combined With Edge Detection and Shape-Based Recognition: Ponomarev *et al.* [30] attempted the segmentation of the femur, by adapting the previously described method (Section V-A5) as follows. After obtaining the binary image grouping the cluster values into one

(a) (b) (c)

Fig. 9. (a) Original image. (b) Preliminary segmented objects. (c) Recognised femur object.

(a) (b) (c)

Fig. 10. Entropy-based segmentation method. (a) Original image. (b) Result after entropy-based segmentation. (c) Final selected femur.

unique value, the method needs to guarantee that only one object is detected as femur. The authors expected the femur bone to have high brightness, large size, contrasted edges, and a central location within the image. These properties were used as features to train a linear support vector machine (SVM) classifier. This was obtained using exhaustive search with 10-fold cross-validation. The whole dataset was divided into 10 parts of equal size. For each iteration, the method was trained on the concatenated set of nine parts and tested on the remaining part. The segmented objects were manually classified into positive and negative classes to train the SVM classifier. This resulted in a scoring function, encoding the recognition model. The femur length was then calculated as the longest distance between any pair of pixels for the selected binary object. An example can be seen in Fig. 9. The parameters required for this method are the coefficients used within the SVM approach. These were adjusted using a cross-validation strategy from the training set.

2) Morphology-Based Approach: Wang *et al.* [31] developed a fully automatic method, based on morphology, to extract the fetal femur bone from the ultrasound images. They proposed two methods for segmenting the femur, one based on entropy and one based on edge detection. The first one was used as the main approach, and the second method was only used when the main approach failed, as an alternative approach.

For the main approach, after the images were initially filtered by a median filter, entropy-based segmentation identified possible pixel candidates within the images, as shown in Fig. 10(b). To obtain the final segmented femurs, first the image complement followed by a morphological dilation were performed for each image. Then, slim and long connected objects can be automatically selected as the final segmented femurs [Fig. 10(c)] by combining the information of density and height-to-width ratio for each segmented object. The density is calculated as the number of segmented pixels over the area of the bounding box for that particular object. The best object is obtained by considering the morphology and layout of the detected objects.

The alternative segmentation approach obtains the horizontal edges and the stretched edges using filters as a preprocessing step. The final step consists of seeking for the longest and slim objects in the resulting edge images. An example is given in Fig. 11.

(a) (b) (c)

Fig. 11. Edge-based segmentation method. (a) Original image. (b) Result after horizontal edges and stretching. (c) Selected femur.

For both methods, the femur length is derived from the segmentation results by using the width and height of the bounding rectangle of the segmented femur object. Two parameters are used within this method: the density of an object and the height to width ratio of an object. These were held constant over all tests.

VI. Experimental Results

In this section, the qualitative and quantitative evaluation for fetal head and fetal femur sub-challenges is presented. All the proposed methods are evaluated against the ground truth on the 90 fetal US images acquired across gestation, as described in Section II-B.

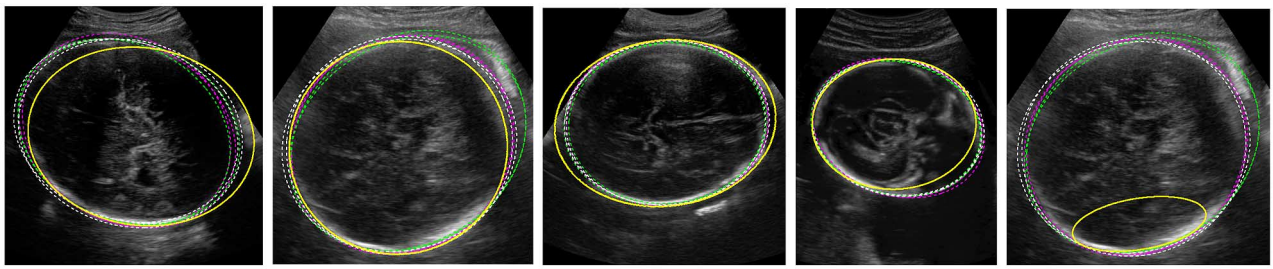
A. Fetal Head Sub-Challenge

1) Failures: No failures were reported for the fetal head sub-challenge, and all the proposed methods obtained segmentation results that overlapped the manually fitted ellipses drawn by the experts.

2) Qualitative Evaluation: Qualitative evaluation was performed on the set of 90 fetal head US images acquired across gestation. The poorest result from each of the proposed methods participating in this challenge is shown in Fig. 12. Note that most of the poor results correspond to images of 33 weeks fetuses, which generally have lower image quality (e.g., increased shadowing due to increased bone density) than earlier gestations. Similarly, the best results, displayed in Fig. 13, were generally at early gestation (21 weeks and 28 weeks), where the image quality is normally better, presenting less artefacts than at later gestation and with clear anatomical definition.

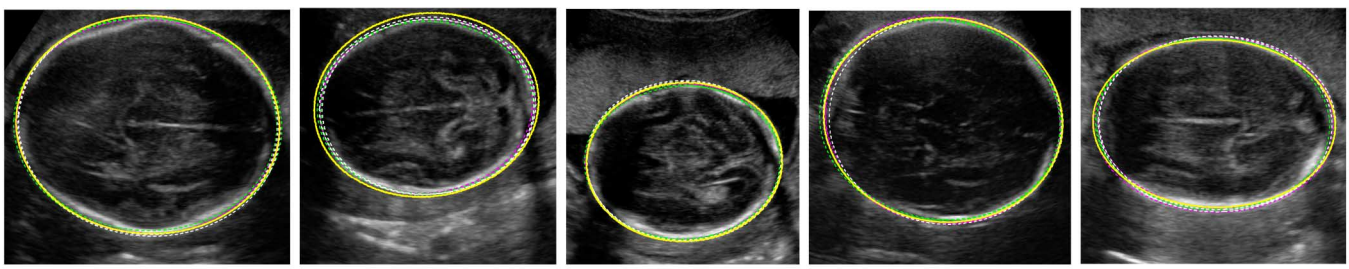
3) Quantitative Evaluation: Table VIII presents the region-based and distance-based evaluation for each proposed method. The best results per metric are highlighted in bold. For the region-based evaluation, the Foi *et al.* method performed best in terms of precision and Dice similarity. Stebbing and McManigle's method performed best in terms of sensitivity. Ciurte *et al.* obtained the best result in terms of specificity. Overall, the Foi *et al.* method had better performance followed closely by Stebbing and McManigle's method.

For the distance-based evaluation, smallest mean error in terms of MSD, ASD, and RMSD is obtained by the Foi *et al.* method, closely followed by Stebbing and McManigle. However, Stebbing and McManigle's method presents a smaller standard deviation, showing that their segmentation is less vari-



(a) Foi et al. 33 wks. (b) Ciurte et al. 33 wks. (c) Stebbing et al. 33 wks. (d) Sun. 21 wks. (e) Ponomarev et al. 33 wks.

Fig. 12. Poorest fetal head result for each proposed method in terms of precision. Yellow continuous lines denote the automatic methods. Dashed lines represent manually fitted ellipses by the clinical experts (magenta: Expert 1, green: Expert 2, white: Expert 3) as defined in Section IV-A.



(a) Foi et al. 28 wks. (b) Ciurte et al. 21 wks. (c) Stebbing et al. 21 wks. (d) Sun. 28 wks. (e) Ponomarev et al. 21 wks.

Fig. 13. Best fetal head result for each proposed method in terms of precision. Yellow continuous lines denote the automatic methods. Dashed lines represent manually fitted ellipses by the clinical experts (magenta: Expert 1, green: Expert 2, white: Expert 3) as defined in Section IV-A.

TABLE VIII
QUANTITATIVE EVALUATION OF THE METHODS FOR THE FETAL HEAD SUB-CHALLENGE

	Region-Based				Distance-Based		
Method	**Precision** (%)	**Sensitivity** (%)	**Specificity** (%)	**Dice** (%)	**MSD** (mm)	**ASD** (mm)	**RMSD** (mm)
Foi et al. [26]	$\mathbf{95.72 \pm 1.92}$	98.51 ± 1.20	98.28 ± 1.26	$\mathbf{97.80 \pm 1.04}$	$\mathbf{2.16 \pm 1.44}$	$\mathbf{0.88 \pm 0.53}$	$\mathbf{1.08 \pm 0.69}$
Ciurte et al. [27]	89.53 ± 2.81	90.19 ± 3.05	$\mathbf{99.62 \pm 0.48}$	94.45 ± 1.57	4.6 ± 1.64	2.10 ± 0.69	2.47 ± 0.83
Stebbing et al. [28]	94.63 ± 1.45	$\mathbf{98.86 \pm 1.26}$	97.53 ± 1.29	97.23 ± 0.77	2.59 ± 1.14	1.07 ± 0.39	1.29 ± 0.51
Sun [29]	94.15 ± 2	95.63 ± 2.46	99.12 ± 1.12	96.97 ± 1.07	3.02 ± 1.55	1.19 ± 0.54	1.48 ± 0.71
Ponomarev et al. [30]	87.29 ± 12.79	88.06 ± 12.88	99.48 ± 1.11	92.53 ± 10.22	6.87 ± 9.82	2.83 ± 3.83	3.55 ± 5.21

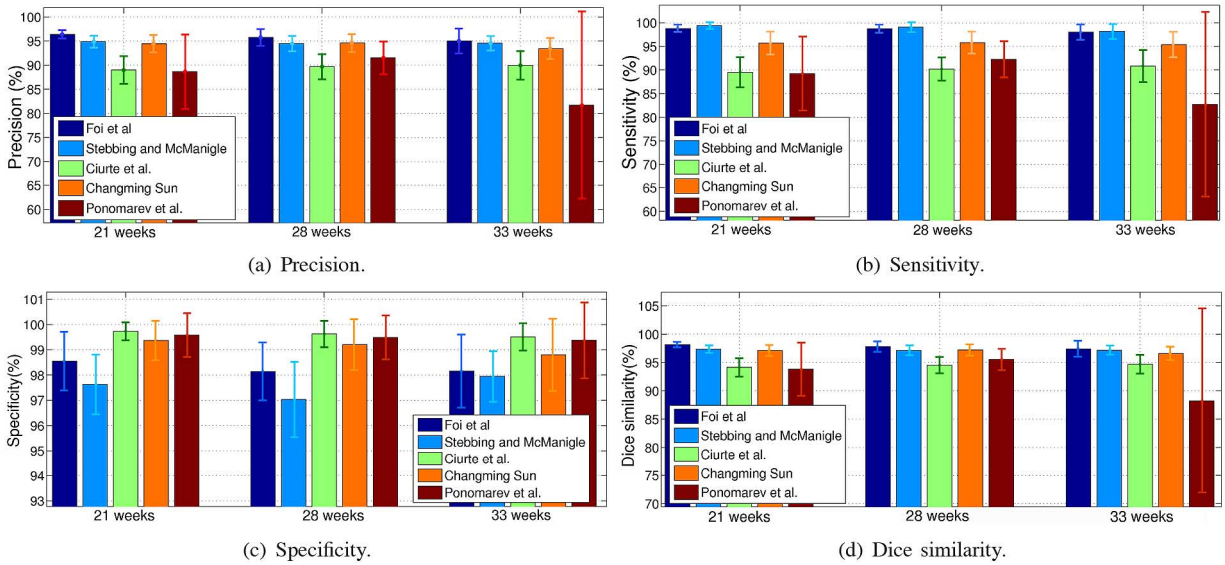
able. Foi *et al.* also obtained similar results to the inter-observer variability presented in Table III, producing results comparable to manual delineation.

To study if the performance varies for the different gestational age groups, the mean and standard deviations in terms of precision, accuracy (sensitivity and specificity), and Dice similarity, at 21, 28, and 33 weeks are presented in Fig. 14. The best performance in terms of mean precision [Fig. 14(a)] for all three gestational ages is by the Foi *et al.* method, closely followed by Stebbing and McManigle's method. However, the Foi *et al.* standard deviation increases marginally across gestation. This might be due to the higher variation in image quality at later gestations and the presence of stronger artefacts. Stebbing and McManigle's method has a small and constant standard deviation across gestation. This is also true for the overall precision presented in Table VIII.

In terms of sensitivity, Stebbing and McManigle and Foi *et al.*'s methods perform better than the other methods according to Fig. 14(b). They also have the smallest standard deviation, which increases slightly at later gestations. Sun's method has a similar performance, with constant mean and standard deviation across gestation. In terms of specificity, all methods seem to have constant means and standard deviations according to Fig. 14(c). The best result is given by Ciurte *et al.* (Table VIII).

In terms of Dice similarity, the Foi *et al.* method had the best result, followed by Stebbing and McManigle and Sun [Fig. 14(d)]. This is also true overall, as shown in Table VIII. Mean and standard deviation appear quite constant for all methods except for the Ponomarev *et al.* method.

The last aspect of the evaluation was to study the performance in terms of clinical measurements derived from the segmented objects. Table IX presents the mean and standard deviation from the Bland–Altman plots in comparison to each expert and over all experts. The best BPD results when compared with the experts were obtained by Sun's method, closely followed by the Foi *et al.* method. The best OFD results were obtained by the Foi *et al.* method, closely followed by Stebbing and McManigle's method. This means that the major axes of the fitted ellipses, from which the OFD measurements are derived, are probably more accurate for the Foi *et al.* method and Stebbing and McManigle's method, whereas the minor axis of the fitted ellipses seems to be better detected by Sun's method. Since the OFD measurement is greater than the BPD measurement, this results



(a) Precision. (b) Sensitivity.

(c) Specificity. (d) Dice similarity.

Fig. 14. Mean and standard deviation for the fetal head for each gestational age in terms of (a) precision; (b) sensitivity; (c) specificity; and (d) Dice similarity.

TABLE IX
BLAND–ALTMAN PLOTS (FETAL HEAD SUB-CHALLENGE): BPD, OFD, AND HC

	Method	Expert 1 (mm)	Expert 2 (mm)	Expert 3 (mm)	All experts (mm)
BPD	**Foi et al. [26]**	-0.94 ± 1.29	-1.15 ± 0.99	$\mathbf{-0.77 \pm 1.11}$	-0.95 ± 1.00
	Ciurte et al. [27]	2.99 ± 1.30	2.78 ± 1.28	3.17 ± 1.32	2.98 ± 1.19
	Stebbing and McManigle [28]	-1.64 ± 1.22	-1.85 ± 0.94	-1.46 ± 1.04	-1.65 ± 0.93
	Sun [29]	$\mathbf{0.59 \pm 1.37}$	$\mathbf{0.38 \pm 1.26}$	0.77 ± 1.43	$\mathbf{0.58 \pm 1.24}$
	Ponomarev et al. [30]	4.69 ± 9.92	4.48 ± 9.92	4.86 ± 9.94	4.67 ± 9.91
OFD	**Foi et al. [26]**	-1.59 ± 2.79	$\mathbf{-0.13 \pm 3.10}$	$\mathbf{-0.48 \pm 2.46}$	$\mathbf{-0.73 \pm 2.52}$
	Ciurte et al. [27]	3.36 ± 3.27	4.81 ± 3.52	4.46 ± 3.12	4.21 ± 3.07
	Stebbing and McManigle [28]	-1.81 ± 3.01	-0.36 ± 3.65	-0.71 ± 2.77	-0.96 ± 2.92
	Sun [29]	$\mathbf{0.59 \pm 3.66}$	2.05 ± 4.04	1.69 ± 3.67	1.45 ± 3.59
	Ponomarev et al. [30]	4.49 ± 7.58	5.95 ± 8.23	5.60 ± 7.17	5.34 ± 7.57
HC	**Foi et al. [26]**	$\mathbf{-1.92 \pm 3.76}$	$\mathbf{-2.67 \pm 4.04}$	$\mathbf{-1.44 \pm 3.52}$	$\mathbf{-2.01 \pm 3.29}$
	Ciurte et al. [27]	12.02 ± 5.60	11.27 ± 5.51	12.51 ± 5.78	11.93 ± 5.32
	Stebbing and McManigle [28]	-3.37 ± 4.44	-4.12 ± 4.77	-2.88 ± 4.14	-3.46 ± 4.06
	Sun [29]	3.92 ± 6.29	3.17 ± 6.05	4.40 ± 5.47	3.83 ± 5.66
	Ponomarev et al. [30]	16.47 ± 24.95	15.72 ± 25.03	16.96 ± 24.85	16.39 ± 24.88

in similar performance of the HC measurement with respect to the OFD, as shown in Table IX.

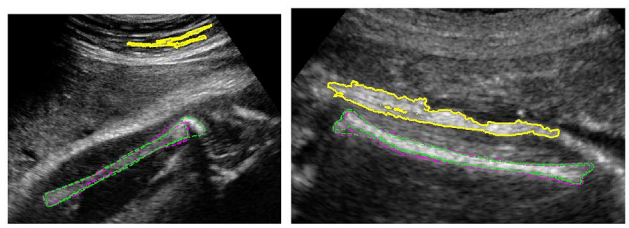
Overall, for the fetal head sub-challenge, the Foi *et al.* method seems to perform best in terms of region-based and distance-based metrics, as well as clinical measurements. Stebbing and McManigle obtained similar results. Sun's method showed high agreement in BPD biometric measurements.

B. Fetal Femur Sub-Challenge

Qualitative and quantitative evaluation is performed in the following for the two methods submitted to the fetal femur US image segmentation challenge. The data set presents different qualities, with some images especially challenging, but all of them used in clinical practice.

1) Failures: The Ponomarev *et al.* method had a total of two failures on different images, shown in Fig. 15. The Wang *et al.* method had a total of four failures over the 90 images in the fetal femur dataset. The failures are presented in Fig. 16.

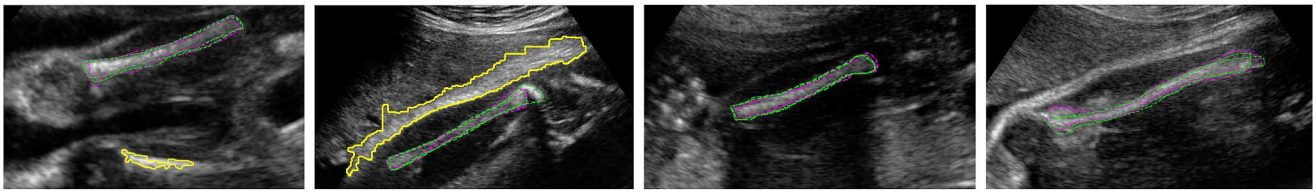
Two of them were due to the method not finding any result on the images. In both cases, the methods found other elongated objects in the images (e.g., other bones, adipose tissue layer, placental tissue) instead of the femur bone. This is because the methods are based on intensities, and the detected incorrect objects had high intensity values while having an elongated shape.



(a) 33 weeks fetus. (b) 33 weeks fetus.

Fig. 15. (a)–(b) Failures of the Ponomarev *et al.* method in terms of precision for the fetal femur. Yellow continuous lines: automatic methods. Dashed lines: manual delineations (magenta: Expert 1, green: Expert 2).

2) Quantitative Evaluation: The evaluation with respect to the measurements is presented in Table X and shows that the best results are obtained by the Wang *et al.* method. Table XI presents the region-based and distance-based evaluation for



(a) 21 weeks fetus. (b) 33 weeks fetus. (c) 21 weeks fetus. (d) 33 weeks fetus.

Fig. 16. (a)–(d) Failures of the Wang *et al.* method in terms of precision for the fetal femur. Yellow continuous lines: automatic methods. Dashed lines: manual delineations (Magenta: Expert 1, Green: Expert 2).

TABLE X
BLAND–ALTMAN PLOTS (FETAL FEMUR SUB-CHALLENGE): FL

Method	**E1** (mm)	**E2** (mm)	**Both** (mm)
Ponomarev et al. [30]	1.80 ± 10.98	3.15 ± 10.91	2.48 ± 10.93
Wang et al. [31]	$\mathbf{1.04 \pm 9.35}$	$\mathbf{2.41 \pm 9.46}$	$\mathbf{1.72 \pm 9.39}$

each proposed method. The best results are highlighted in bold. In terms of region-based metrics, the Ponomarev *et al.* method seems to have a higher performance, whereas the Wang *et al.* method obtains better results in terms of the distance-based evaluation. However, unlike the fetal head challenge, the measurement results are inferior to those obtained manually between experts (cf. Table VI) with much higher variability.

3) Qualitative Evaluation: Qualitative evaluation was performed on the fetal femur data set of 90 ultrasound images acquired at three different gestational ages (21, 28, and 33 weeks). The poorest and best results obtained from each method are shown in Fig. 17. Notice how the poorest results [Fig. 17(a) and (c)] are only segmenting the brightest part of the femur. This is due to the high inhomogeneities existing within the femur. The best results [Fig. 17(b) and (d)] perform similarly to manual delineations.

C. Efficiency

Since the algorithms have been programmed using different software, computers, and programming languages, the study of efficiency cannot thoroughly be performed. Even if efficiency was not one of the aims of this challenge, and that some methods are more expensive computationally than others, we report on the times and specifications used in the presented implementations as shown in Table XII. After the challenge, some of the teams optimized their code and were able to reduce these times considerably.

VII. DISCUSSION

A. Fetal Head Sub-Challenge

The five methods submitted to the fetal head sub-challenge are very different and focus on either image appearance or edge information. The methods of Ciurte *et al.* [27] and Sun [29] used graph-based approaches. The Foi *et al.* method [26] was based on signal processing combined with an optimization framework. Stebbing and McManigle [28] used a machine learning approach based on a boundary fragment model resulting from a training step. Four out of five methods obtained constant results across gestation [cf. Fig. 14(a)–(d)] except the Ponomarev *et al.* method [30], which got variable means and standard deviations for the three gestational age groups, the poorest results being at 33 weeks, where the images have in general lower quality and present more artefacts. Since the Ponomarev *et al.* method uses the appearance of the object of interest to define the multiple thresholds, and then fits an ellipse to the resulting binary image, it is to be expected that the results are more linked to the image quality than the other methods, which relied less on the appearance of the object of interest.

The Foi *et al.* method obtained the best results overall, achieving a mean and standard deviation close to the ground truth values for both region-based and distance-based metrics (Table VIII), showing a performance as good as the inter-observer variability (Table III).

Stebbing and McManigle's method obtained results close to the Foi *et al.* method, and most of the time had smaller constant standard deviations across gestational ages, which indicates that their method was slightly more consistent.

Sun's method produced results ranked third overall by using a graph-based approach. His method seems also robust across gestational age groups with a high mean and small standard deviation. He obtained the best results for BPD measurements compared to the ground truth (Table IX). Considering that the BPD value is derived from the small axis of the fitted ellipse, this suggests that his method fitted the ellipses better in the small axis direction.

The Ciurte *et al.* method obtained slightly worse results than Foi *et al.*, Stebbing and McManigle, and Sun's methods. It was noted during the workshop that their method was finding the inner contour of the skull instead of the outer contour, which could be the cause of the difference between the other methods. This behavior can be appreciated in Fig. 12(b) and Fig. 13(b). The other methods were finding the outer edge. The Ciurte *et al.* method had constant mean and standard deviation across gestational age groups [cf. Fig. 14(a)–(d)], with a consistent performance for different image qualities. It may be that, if their method was modified to detect the outer contours of the skull, the results would have improved and may have been comparable to the other methods that performed better.

In terms of reproducibility of clinical measurements (Table IX), only the method by Ponomarev *et al.* had a much higher standard deviation than the inter-expert variability presented in Table V in all cases. This shows that this method had a much lower reproducibility than manual delineations. For the BPD measurement, all the other methods had lower standard deviation than the inter-expert variability. For the

TABLE XI
QUANTITATIVE EVALUATION OF THE METHODS FOR THE FETAL FEMUR SUB-CHALLENGE

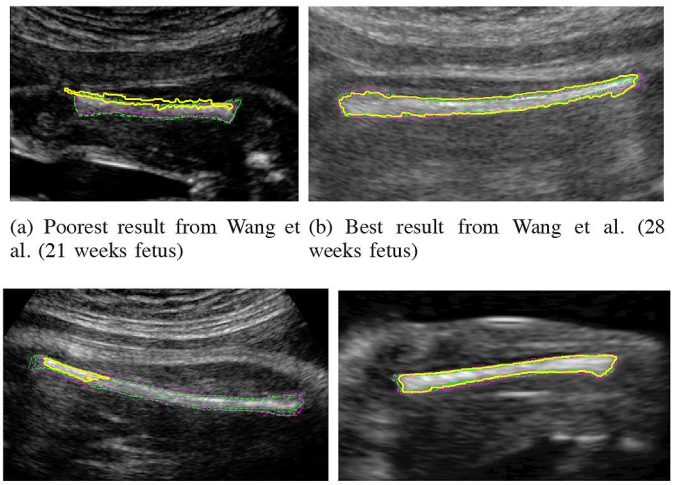
	Region-Based				Distance-Based		
Method	**Precision** (%)	**Sensitivity** (%)	**Specificity** (%)	**Dice** (%)	**MSD** (mm)	**ASD** (mm)	**RMSD** (mm)
Ponomarev et al. [30]	$\mathbf{65.44 \pm 16.98}$	$\mathbf{72.79 \pm 19.40}$	$\mathbf{99.70 \pm 0.39}$	$\mathbf{77.40 \pm 15.35}$	6.39 ± 9.53	1.23 ± 2.3	2.04 ± 3.76
Wang et al. [31]	60.56 ± 15.88	69.84 ± 20.36	99.66 ± 0.37	73.95 ± 14.56	$\mathbf{6.02 \pm 7.29}$	$\mathbf{1.04 \pm 1.29}$	$\mathbf{1.77 \pm 2.41}$

OFD measurement, only the method by Foi *et al.* had a reproducibility within the range reported in Table V. The other methods had a slightly higher standard deviation, but close to the ground truth values, except for the Ponomarev *et al.* method. For the HC measurement, only the methods by Foi *et al.* and Stebbing and McManigle had a reproducibility close to manual segmentations. The other methods, except the one by Ponomarev *et al.*, obtained values within the range reported in previous reproducibility studies [41], [42].

B. Fetal Femur Sub-Challenge

Fetal femur segmentation is the harder of the two sub-challenges. The complete segmentation of the femur needs to take into account the huge inhomogeneities existing in the object of interest. The two methods participating in this challenge relied on appearance and edge information extracted directly from intensity values. This makes the methods rely on the image quality and in some cases miss certain parts of the femur bone during the segmentation process. It also makes the methods more prone to finding other objects that are not the femur bone but have similar appearance to it, hence producing failures on some of the images. This challenge would require a more advanced modelling of the femur bone, incorporating morphological measures of normal fetal femur across gestational ages. The overall mean values in terms of precision, sensitivity, and Dice similarity, were lower than those obtained manually (Table XI). The overall standard deviations of both methods ranged around 16%–17% for precision, 19%–20% for sensitivity, and 14%–15% for Dice similarity, indicating a high variability with respect to the ground truth. Manual segmentations presented a standard deviation around 5% for precision, 6% sensitivity, and 3% Dice similarity for intra-expert variability (Table VI). Considering the inter-expert variability, precision had a standard deviation around 6%, sensitivity around 7%, and Dice similarity around 4%.

This sub-challenge proves the necessity of using both region and distance-based metrics for segmentation evaluation. The Ponomarev *et al.* method [30] obtained better results in terms of precision, specificity, sensitivity, and Dice similarity, whereas the Wang *et al.* method [31] achieved better results in terms of MSD, ASD, and RMSD values (Table XI). The Ponomarev *et al.* method had higher overlap with respect to manual segmentations than the Wang *et al.* method, and slightly higher distance-based errors. This could also be due to the fact that the Wang *et al.* method had two more failures than the Ponomarev *et al.* method, and the evaluation results were reported only in the images where there were no failures. Therefore, the Ponomarev *et al.* had slightly higher MSD, ASD, and RMSD, but these were calculated on two more images than in the Wang *et al.* method. In terms of the actual FL measurement, the Wang *et al.* method obtained better results for both mean and standard deviation (Table X). However, in terms of reproducibility of the FL measurement (cf. Table X), none of the methods obtained a standard deviation within the values reported in Table VII, which means that they had low reproducibility compared to manual delineations.



(a) Poorest result from Wang et al. (21 weeks fetus) (b) Best result from Wang et al. (28 weeks fetus)

(c) Poorest result from Ponomarev et al. (33 weeks fetus) (d) Best result from Ponomarev et al. (21 weeks fetus)

Fig. 17. Poorest and best fetal femur results for each proposed method in terms of precision. Yellow continuous lines: automatic methods. Dashed lines: manual delineations (magenta: Expert 1, green: Expert 2).

Another discrepancy to note was that some of the Wang *et al.* segmentation results only found the brightest part of the bone, resulting in an incomplete segmentation result, but a good FL measurement could still be observed [e.g., Fig. 17(a)]. When the femur's appearance had homogeneous intensity values, both methods seemed to perform well, agreeing with manual delineations [e.g., Fig. 17(b) and (d)].

C. General Observations

Signal processing methods, graph-based methods, and machine learning methods seemed to achieve a good performance, since they considered the images as a whole. Some of these also take into account the relationship between different regions of the images simultaneously. On the contrary, intensity and gradient-based methods have a lower performance, since they rely more on the appearance of the objects of interest, which present high variability.

D. How to Move the Fetal us Image Segmentation Field Forward?

From the four sub-challenges proposed, only the head and femur challenges were attempted. The abdomen and whole

TABLE XII
COMPUTER SPECIFICATIONS AND EFFICIENCY

Method	Time per image	Computer Specifications	OS	Software
Foi et al. [26]	5 min	PC Intel Core 2 Duo 2.6GHz, 8GB (Laptop, one core)	Windows 7 64-bit	Matlab
Ciurte et al. [27]	196s	PC Intel Core 2 Duo 2.66GHz, 2GB	Windows 7	Matlab
Stebbing and McManigle [28]	100s-150s	Intel Pentium Core 2 Duo 3.40GHz (one core)	Linux Fedora 13	Python C++
Sun [29]	1.5s	PC Intel Core 2 Duo 3GHz, 2GB	Linux	C
Ponomarev et al. [30]	19.6s (Head) 24.2s (Femur)	MacBook Pro, 2.53 GHz Intel Core i5	Mac OS X	C++
Wang et al. [31]	2.28s	Laptop Intel Core i7 2.8GHz, 8GB	Windows 7	C#

fetus segmentations are extremely challenging due to the lack of strong object boundaries and the similar appearance of surrounding objects. General frameworks that could solve all four sub-challenges simultaneously are yet to be developed. As argued in [6], successful US image segmentation methods are normally application dependent.

One of the main difficulties of working with US images is that they have a variable appearance and it is difficult to obtain quantitative measures of quality to provide more insight on the performance of different methods with respect to image quality in an objective manner. We need better tools of quantitatively assessing US image quality, to be able to study method's performance in depth, relating the performance to the quality of the image. This can be extrapolated to other imaging modalities, but it is especially important in a modality like US. This is a problem that is not solved yet and needs further investigation.

VIII. CONCLUSION

This paper presented a thorough qualitative and quantitative segmentation evaluation of the representative selection of current methods submitted to *Challenge US: Biometric Measurements from Fetal Ultrasound Images*, held at ISBI 2012. The images were selected to incorporate the different qualities, reflective of a real antenatal clinical environment. Three different gestational ages were assessed to incorporate image variability across gestation. Several experts manually delineated the objects of interest to define the ground truth, which was used within the evaluation framework. A total of five teams submitted their results to the fetal head sub-challenge and two teams to the fetal femur sub-challenge, including one team who attempted both.

The results for the fetal head sub-challenge show that a very good performance can be achieved and that it is comparable to manual delineations. Several groups produced results that could be potentially used in clinical settings. The fetal femur sub-challenge consisted of solving a very hard segmentation problem, since the object of interest has strong appearance changes within the object. Furthermore, other elongated objects are present around the femur bone, causing methods to fail in certain situations. The performance of the femur sub-challenge was inferior to the head sub-challenge, because the task was more complex and the techniques used relied more on the femur's appearance.

Further investigation is necessary to provide better quantitative tools for assessing US image quality, which in turn will assist in developing a better understanding of how images cope with the image quality variability.

On release of the data, anticipated autumn 2014, the website (http://www.ibme.ox.ac.uk/challengeus2012) will provide a mechanism to upload new segmentation results and compare them to previous methods.

APPENDIX A
IMAGE QUALITY SCORING CRITERIA

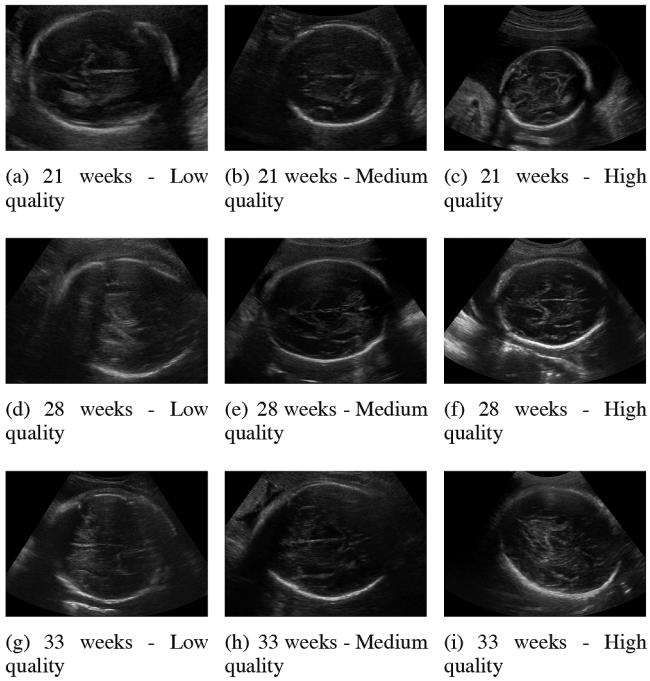
In the following, we present the scoring criteria used for selecting images of different qualities within each sub-challenge. The scoring criteria is based on a score-based method for quality control [51]–[53]. The aim was to select as many good, medium, and high quality images within each gestational age group as objectively as possible. Considerations of image quality are not independent of the gestational ages considered, as they describe quality variation of cross-sectional data (data in a certain gestational age window), which capture a wide range of image quality factors. The scoring criteria was different for each sub-challenge and was performed by experts on each type of images, taking into account the image characteristics. In the case of fetal ultrasound images, the fetal anatomy varies during pregnancy, as well as soft tissue properties and composition. The quality of the images diminishes with gestational age as shown in Figs. 18 and 19 for the head and femur, respectively. This is due to the increase of the body mass index of the mother towards the end of pregnancy, the increase in fetal bone density and fetal size, the reduced amniotic fluid present in older fetuses, and the changes of tissue texture, which create different speckle patterns at different gestational ages.

The scoring criteria used to classify the fetal head images (Fig. 18) between low, medium, and high score is as follows.

Low: Skull is not symmetrical or elliptical in shape. Skull boundary is barely visible. Internal anatomy is difficult to discern with an overall lack of contrast in the image.

Medium: The fetal skull should be roughly elliptical in shape. Skull boundary visible but not less than 60% encirclement. Rough internal anatomy visible (lateral ventricles, cerebral falx, cavum septum pellucidum, thalamus). Average contrasted image.

High: Fetal skull should be roughly symmetrical and elliptical in shape. Skull boundary should be visible with more

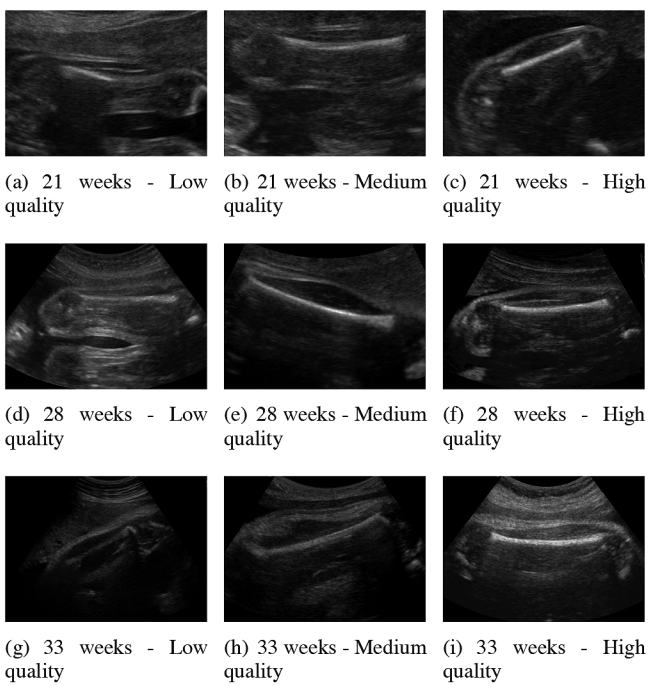


(a) 21 weeks - Low quality (b) 21 weeks - Medium quality (c) 21 weeks - High quality

(d) 28 weeks - Low quality (e) 28 weeks - Medium quality (f) 28 weeks - High quality

(g) 33 weeks - Low quality (h) 33 weeks - Medium quality (i) 33 weeks - High quality

Fig. 18. Ultrasound images of the head at (a)–(c) 21 weeks of gestation, (d)–(f) 28 weeks of gestation, and (g)–(i) 33 weeks of gestation.



(a) 21 weeks - Low quality (b) 21 weeks - Medium quality (c) 21 weeks - High quality

(d) 28 weeks - Low quality (e) 28 weeks - Medium quality (f) 28 weeks - High quality

(g) 33 weeks - Low quality (h) 33 weeks - Medium quality (i) 33 weeks - High quality

Fig. 19. Ultrasound images of the femur at (a)–(c) 21 weeks of gestation, (d)–(f) 28 weeks of gestation, and (g)–(i) 33 weeks of gestation.

than 60% encircled cranial area. Internal anatomy (lateral ventricles, cerebral falx, cavum septum pellucidum, thalamus, and cortical boundary) visible and discernible. High contrasted image.

The scoring criteria used to classify the fetal femur images (Fig. 19) considered.

- The sharpness of the border of the femur from all directions.
- How easy it is to find femur end points (distal and proximal).
- The difference between the femur and its surrounding tissues. Better femur images need to have bright femur and relatively dark surrounding.
- The continuity of the femur. Some femurs are not fully visible because of the scan signal direction and shadowing so these are low quality.

We can observe that more artefacts appear in the images towards the end of pregnancy as a result of the fetus becoming bigger and compressed within the womb, with less space to move. The bone density in the fetus increases too, creating shadows and splaying in the ultrasound images. These shadows appear in the skull in the head, in the ribs and spine in the abdomen, and in the femur in the leg, respectively. Changes in size, shape, pose, and composition are also important, especially in the abdomen that is a soft body region in comparison to the bony structures of head and femur.

ACKNOWLEDGMENT

S. Rueda, J. A. Noble, and C. L. Knight acknowledge the Wellcome/EPSRC Centre of Excellence in Medical Engineering—Personalised Healthcare, WT 088877/Z/09/Z. The work of A. T. Papageorghiou was supported by the Oxford Partnership Comprehensive Biomedical Research Centre funded by the Department of Health NIHR Biomedical Research Centres funding scheme. The work of M. Yaqub was supported by the EPSRC Grant EP/G030693/1. The work of A. Foi, M. Maggioni, A. Pepe, and J. Tohka was supported in part by the Academy of Finland under Grant 130275 and Grant 252547, in part by the Tampere Graduate School in Information Science and Engineering (TISE), and in part by the Finnish Doctoral Programme in Computational Sciences (FICS). The work of A. Ciurte, X. Bresson, M. Bach Cuadra was supported by the Center for Biomedical Imaging (CIBM) of the Geneva-Lausanne Universities and the EPFL, as well as the foundations Leenaards and Louis-Jeantet. The work of R.V. Stebbing and J. E. McManigle's was supported by the Rhodes Trust, in part by the U.S. National Institutes of Health (NIH) Graduate Partnership Program (GPP), and in part by the National Heart, Lung, and Blood Institute (NHLBI) intramural research program. The work of G. V. Ponomarev, M. S. Gelfand, and M. D. Kazanov was supported by the Ministry of Education and Science of Russian Federation under Project 8049. The medical images were provided by the International Fetal and Newborn Growth Consortium, INTERGROWTH-21st, Nuffield Department of Obstetrics and Gynaecology, John Radcliffe Hospital, University of Oxford, Oxford, U.K. The authors would like to thank Prof. J. K. Udupa for the useful discussions on segmentation evaluation.

REFERENCES

[1] P. Loughna, L. Chitty, T. Evans, and T. Chudleigh, "Fetal size and dating: Charts recommended for clinical obstetric practice," *Ultrasound*, vol. 17, no. 3, pp. 161–167, 2009.

[2] B. Hearn-Stebbins, "Normal fetal growth assessment: A review of literature and current practice," *J. Diagn. Med. Sonog.*, vol. 11, no. 4, pp. 176–187, July 1995.

[3] M. Pramanik, M. Gupta, and K. B. Krishnan, "Enhancing reproducibility of ultrasonic measurements by new users," *Proc. SPIE Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*, vol. 8673, p. 86730Q, 2013.

[4] G. Carneiro, B. Georgescu, and S. Good, "Knowledge-based automated fetal biometrics using syngo Auto OB measurements," *Siemens Medical Solutions*, 2008.

[5] J. Espinoza, S. Good, E. Russell, and W. Lee, "Does the use of automated fetal biometry improve clinical work flow efficiency?," *Journal of Ultrasound in Medicine*, vol. 32, no. 5, pp. 847–850, 2013.

[6] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: A survey," *IEEE Trans. Med. Imaging.*, vol. 25, no. 8, pp. 987–1010, 2006.

[7] J. G. Thomas, R. A. Peters, and P. Jeanty, "Automatic segmentation of ultrasound images using morphological operators," *IEEE Trans. Med. Imaging*, vol. 10, no. 2, pp. 180–186, 1991.

[8] J. G. Thomas, P. Jeanty, R. A. Peters, and E. A. Parrish, "Automatic measurements of fetal long bones. a feasibility study," *J. Ultrasound Med.*, vol. 10, no. 7, pp. 381–385, July 1991.

[9] B. Rahmatullah and R. Besar, "Analysis of semi-automated method for femur length measurement from foetal ultrasound," *J. Med. Eng. Technol.*, vol. 33, no. 6, pp. 417–425, 2009.

[10] V. Shrimali, R. S. Anand, and V. Kumar, "Improved segmentation of ultrasound images for fetal biometry, using morphological operators," *IEEE Eng. Med. Biol. Soc.*, vol. 2009, pp. 459–462, 2009.

[11] C. W. Hanna and A. B. M. Youssef, "Automated measurements in obstetric ultrasound images," in *ICIP*, Oct. 1997, pp. 504–507.

[12] W. Lu and J. Tan, "Segmentation of ultrasound fetal images," *Biological Quality and Precision Agriculture II*, vol. 4203, no. 1, pp. 81–90, 2000.

[13] W. Lu, J. Tan, and R. Floyd, "Automated fetal head detection and measurement in ultrasound images by iterative randomized Hough transform," *Ultrasound Med. Biol.*, vol. 31, no. 7, pp. 929–936, July 2005.

[14] V. Chalana, T. C. Winter, D. R. Cyr, D. R. Haynor, and Y. Kim, "Automatic fetal head measurements from sonographic images," *Acad. Radiol.*, vol. 3, no. 8, pp. 628–635, Aug. 1996.

[15] S. D. Pathak, V. Chalana, and Y. Kim, "Interactive automatic fetal head measurements from ultrasound images using multimedia computer technology," *Ultrasound Med. Biol.*, vol. 23, no. 5, pp. 665–673, 1997.

[16] S. D. Pathak, V. Chalana, and Y. Kim, "Multimedia systems in ultrasound image boundary detection and measurements," *Proceedings of SPIE Medical Imaging 1997: Image Display*, vol. 3031, pp. 397–408, 1997.

[17] S. M. G. V. B. Jardim and M. A. T. Figueiredo, "Automatic contour estimation in fetal ultrasound images," in *ICIP*, Sept. 2003, vol. 2–3, pp. II-1065–1068.

[18] S. M. G. V. B. Jardim and M. A. T. Figueiredo, "Segmentation of fetal ultrasound images," *Ultrasound Med. Biol.*, vol. 31, no. 2, pp. 243–250, Feb. 2005.

[19] B. P. Shan and M. Madheswaran, "Extraction of fetal biometrics using class separable shape sensitive approach for gestational age estimation," in *ICCTD*, 2009, pp. 376–380.

[20] J. Nithya and M. Madheswaran, "Detection of intrauterine growth retardation using fetal abdominal circumference," *ICCTD*, pp. 371–375, 2009.

[21] J. Yu, Y. Wang, P. Chen, and Y. Shen, "Fetal abdominal contour extraction and measurement in ultrasound images," *Ultrasound Med. Biol.*, vol. 34, no. 2, pp. 169–182, Feb. 2008.

[22] G. Carneiro, B. Georgescu, S. Good, and D. Comaniciu, "Automatic fetal measurements in ultrasound using constrained probabilistic boosting tree," in *MICCAI*, 2007, vol. 10, pp. 571–579, Part II, LNCS 4792.

[23] G. Carneiro, B. Georgescu, S. Good, and D. Comaniciu, "Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree," *IEEE Trans Med. Imaging*, vol. 27, no. 9, pp. 1342–1355, Sept. 2008.

[24] J. Yu, Y. Wang, and P. Chen, "Fetal ultrasound image segmentation system and its use in fetal weight estimation," *Med. Biol. Eng. Comput.*, vol. 46, no. 12, pp. 1227–1237, Dec. 2008.

[25] International Fetal and Newborn Growth Consortium, The International Fetal and Newborn Growth Standards for the 21st Century (Intergrowth-21st) Study Protocol [Online]. Available: www.intergrowth21.org.uk 2008

[26] A. Foi, M. Maggioni, A. Pepe, and J. Tohka, "Head contour extraction from the fetal ultrasound images by difference of gaussians revolved along elliptical paths," in *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, 2012, pp. 1–3.

[27] A. Ciurte, X. Bresson, and M. B. Cuadra, "A semi-supervised patch-based approach for segmentation of fetal ultrasound imaging," in *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, 2012, pp. 5–7.

[28] R. V. Stebbing and J. E. McManigle, "A boundary fragment model for head segmentation in fetal ultrasound," in *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, 2012, pp. 9–11.

[29] C. Sun, "Automatic fetal head measurements from ultrasound images using circular shortest paths," in *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, 2012, pp. 13–15.

[30] G. V. Ponomarev, M. S. Gelfand, and M. D. Kazanov, "A multilevel thresholding combined with edge detection and shape-based recognition for segmentation of fetal ultrasound images," in *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, 2012, pp. 17–19.

[31] C.-W. Wang, H.-C. Chen, C.-W. Peng, and C.-M. Hung, "Automatic femur segmentation and length measurement from fetal ultrasound images," *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, pp. 21–23, 2012.

[32] C. H. Chang, P. Y. Tsai, C. H. Yu, H. C. Ko, and F. M. Chang, "Prenatal detection of fetal growth restriction by fetal femur volume: efficacy assessment using three-dimensional ultrasound," *Ultrasound Med. Biol.*, vol. 33, no. 3, pp. 335–341, 2007.

[33] P. A. Mahon, C. Cooper, S. R. Crozier, and K. M. Godfrey, "The use of 3d ultrasound to investigate fetal bone development," *Norsk Epidemiologi*, vol. 19, no. 1, pp. 45–52, 2009.

[34] M. Yaqub, M. K. Javaid, C. Cooper, and J. A. Noble, "Improving the classification accuracy of the classic rf method by intelligent feature selection and weighted voting of trees with application to medical image segmentation," in *Workshop on Machine Learning in Medical Imaging (MLMI)—MICCAI*, 2011, vol. 7009, pp. 184–192.

[35] C. Ioannou, M. Yaqub, A. Noble, K. Javaid, and A. Papageorghiou, "Fetal femur volume measurement using fusion of multiple three-dimensional ultrasound image sets: A pilot study," *Ultrasound Obstet. Gynecol.*, vol. 36, no. S1, p. 80, 2010.

[36] J. K. Udupa, V. R. Leblanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch, and J. Woodburn, "A framework for evaluating image segmentation algorithms," *Comput. Med. Imaging Graph*, vol. 30, no. 2, pp. 75–87, 2006.

[37] T. Heimann, B. van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. M. M. Cashman, Y. Chi, A. Cordova, B. M. Dawant, M. Fidrich, J. D. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmller, R. I. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H.-P. Meinzer, G. Nemeth, D. S. Raicu, A.-M. Rau, E. M. van Rikxoort, M. Rousson, L. Rusko, K. A. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. M. Waite, A. Wimmer, and I. Wolf, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imaging*, vol. 28, no. 8, pp. 1251–1265, Aug. 2009.

[38] D. Huttenlocher, D. Klanderman, and A. Rucklige, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, 1993.

[39] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 1, pp. 307–310, 1986.

[40] J. M. Bland and D. G. Altman, "Measurement error," *BMJ*, vol. 313, pp. 744–753, 1996.

[41] S. Perni, F. A. Chervenak, R. B. Kalish, S. Magherini-Rothe, M. Predanic, J. Streltzoff, and D. W. Skupski, "Intraobserver and interobserver reproducibility of fetal biometry," *Ultrasound Obstet. Gynecol.*, vol. 24, no. 6, pp. 654–658, 2004.

[42] I. Sarris, C. Ioannou, P. Chamberlain, E. Ohuma, F. Roseman, L. Hoch, D. G. Altman, and A. T. Papageorghiou, "International fetal and newborn consortium for the 21st century (INTERGROWTH-21st intra- and interobserver variability in fetal ultrasound measurements," *Ultrasound Obstet Gynecol*, vol. 39, no. 3, pp. 266–273, 2012.

[43] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[44] A. Ciurte, N. Houhou, S. Nedevschi, A. Pica, F. L. M. J.-P. Thiran, X. Bresson, and M. B. Cuadra, "An efficient segmentation method for ultrasound images based on a semi-supervised approach and patch-based features," in *ISBI*, 2011, pp. 969–972.

[45] N. Houhou, X. Bresson, A. Szlam, T. F. Chan, and J.-P. Thiran, "Semi-supervised segmentation based on non-local continuous min-cut," in *SSVM*, 2009, pp. 112–123.

[46] J. Flusser, "On the independence of rotation moment invariants," *Pattern Recogn*, vol. 33, no. 9, pp. 1405–1410, Oct. 2000.

[47] R. V. Stebbing, J. E. McManigle, and J. A. Noble, "Interpreting edge information for improved endocardium delineation in echocardiograms," in *ISBI*, 2012, pp. 238–241.

[48] P. Kovesi, "Image features from phase congruency," *Videre*, vol. 1, no. 3, 1999.

[49] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1270–1281, 2008.

[50] B. Appleton and C. Sun, "Circular shortest paths by branch and bound," *Pattern Recogn.*, vol. 36, no. 11, pp. 2513–2520, 2003.

[51] L. J. Salomon, J. P. Bernard, M. Duyme, B. Doris, N. Mas, and Y. Ville, "Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester," *Ultrasound Obstet. Gynecol.*, vol. 27, no. 1, pp. 34–40, 2006.

[52] L. J. Salomon, N. Winer, J. P. Bernard, and Y. Ville, "A score-based method for quality control of fetal images at routine second-trimester ultrasound examination," *Prenat. Diagn.*, vol. 28, no. 9, pp. 822–827, 2008.

[53] L. J. Salomon, M. Nassar, J. P. Bernard, Y. Ville, A. Fauconnier, and S. F. Pour, "A score-based method to improve the quality of emergency gynaecological ultrasound examination," *Eur. J Obstet. Gynecol. Reprod. Biol.*, vol. 143, no. 2, pp. 116–120, 2009, L'Amélioration des Pratiques Echographiques (SFAPE).