

# Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

## **Learning to segment key clinical anatomical structures in fetal neurosonography informed by a region-based descriptor**

Ruobing Huang  
Ana Namburete  
Alison Noble

# Learning to segment key clinical anatomical structures in fetal neurosonography informed by a region-based descriptor

Ruobing Huang,\* Ana Namburete, and Alison Noble

University of Oxford, Institute of Biomedical Engineering, Department of Engineering Science, Oxford, United Kingdom

**Abstract.** We present a general framework for automatic segmentation of fetal brain structures in ultrasound images inspired by recent advances in machine learning. The approach is based on a region descriptor that characterizes the shape and local intensity context of different neurological structures without explicit models. To validate our framework, we present experiments to segment two fetal brain structures of clinical importance that have quite different ultrasonic appearances—the corpus callosum (CC) and the choroid plexus (CP). Results demonstrate that our approach achieves high region segmentation accuracy (dice coefficient:  $0.81\% \pm 0.06$  CC,  $0.76\% \pm 0.08$  CP) relative to human delineation, whereas the derived automated biometry measurement deviations are within human intra/interobserver variations. The use of our proposed method may help to standardize intracranial anatomy measurements for both the routine examination and the detection of congenital conditions in the future. © 2018 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.5.1.014007](https://doi.org/10.1117/1.JMI.5.1.014007)]

Keywords: segmentation; fetal neurosonography; corpus callosum; choroid plexus.

Paper 17271R received Sep. 15, 2017; accepted for publication Feb. 13, 2018; published online Mar. 10, 2018.

## 1 Introduction

Automated fetal neurosonography is an area of growing interest in medical image analysis to support studies of the developing brain. Clinically, prenatal diagnosis of brain abnormalities using ultrasound (US) has been found to be highly consistent with autopsy findings and fetal magnetic resonance imaging (MRI).<sup>1,2</sup> The fetal anomaly US scan conducted at around 20 gestation weeks requires, typically, three planes of the fetal brain to be acquired to assess growth by biometry of the skull (e.g., head circumference or biparietal diameter measurement), to provide a basic check on neurological development of key brain structures. A full fetal neurosonography scan is more comprehensive (for instance, as described in clinical guidelines<sup>3</sup>) and requires careful capture of a number of standard image views. This is done only when there is suspicion of neurological conditions or damage and, in practice, its use is limited by availability of qualified examiners able to interpret the relatively complex brain anatomy observed in US.

Some clinical studies have reported the size of anatomies such as the atrium [which contains the choroid plexus (CP)] and the cerebellum, and have related these to fetal health.<sup>4,5</sup> However, as these brain structures can be difficult to identify, measurement of them is not included in sonographic examination standards. This is particularly true in late gestation when skull bone begins to calcify. To address this issue, we propose an automatic segmentation framework of brain structures for fetal neurosonography, which also generates biometry measurements simultaneously.

## 1.1 Related Work

Prior work in medical image analysis has considered applying machine learning to automatic detection of fetal brain structures, rather than segmentation. For instance, Carneiro et al.<sup>6</sup> presented a detection method using a probabilistic model, whereby the posterior classifiers were trained by probabilistic boosting trees. Namburete et al.<sup>7</sup> proposed a method to detect the presence of the CP in two-dimensional (2-D) US using AdaBoost with statistical image representations estimated by the Nakagami distribution. Sofka et al.<sup>8</sup> described a technique to detect several fetal brain structures using sequential sampling, with the posterior and transition distributions derived as Gaussian distributions.

Relative to object detection, segmentation provides additional information about a structure's shape and appearance. Skull segmentation in fetal US has attracted most attention as a precursor to head biometry. Earlier work used morphological operations and the Hough transform.<sup>9–11</sup> Such approaches are sensitive to the quality of image acquisition and none of the early methods were validated on large datasets. To the best of our knowledge, there are only two prior publications on US segmentation of fetal brain structures. Yaqub et al.<sup>12</sup> reported a semiautomatic approach to segment fetal brain structures in three-dimensional (3-D) US, where they restricted the search region to the smallest cuboid that enclosed the structure. Cuboid extraction was based on prior knowledge of the approximate location and the size of the structure of interest. Gutiérrez-Becker et al.<sup>13</sup> presented an algorithm to segment the fetal brain cerebellum using a statistic shape model. While they reported promising results, validation was limited to only 20 subjects (the number of volumes they used was not clear in the paper) and it is unclear how that method could be readily

\*Address all correspondence to: Ruobing Huang, E-mail: [ruobing.huang@eng.ox.ac.uk](mailto:ruobing.huang@eng.ox.ac.uk)

extended to other types of brain structures. Furthermore, all images were acquired to a relatively strict protocol leading to the cerebellum appearing in a similar initial position for all the images. This suggests that the solution is not well suited for general clinical imaging practice, where the initial position of brain anatomy is not consistent.

In MRI studies, approaches such as statistical atlas models<sup>14–16</sup> and patch-based techniques<sup>17–19</sup> have been proposed for fetal/adult brain segmentation. Nevertheless, importantly the appearance of structures in MRI images is quite different from the US case. Fetal sonography does not lend itself well to anatomical atlas-based techniques as anatomical structures have varied appearance depending on object tissue properties (echogenicity), and the pose of the object with respect to the transducer (fetal motion makes this a practical challenge to control). Appearance of anatomical boundaries is typically incomplete and strong acoustic shadows can obscure structures. An US-specific approach is thus needed to tackle the challenges of brain structure segmentation in fetal neurosonography.

## 1.2 Proposed Framework

In this paper, we present an approach to automate US segmentation of fetal brain structures. The originality of our work lies in transforming the complex structure delineation problem into a region classification task that utilizes a region-based descriptor. We report experiments on segmenting two structures that are important in monitoring in the developing brain, the corpus callosum (CC) and the CP, to prove both the usefulness of the approach and its generality.

Figure 1 shows an overview of the automated segmentation framework (details of the method are given in Sec. 2). Our proposed approach consists of two steps: (1) candidate region selection and (2) region characterization using the proposed region descriptor combined with a boosting classifier trained to identify the desired candidate region. In testing, an unseen image is passed through the same pipeline to generate an automated segmentation.

The outline of the article is as follows: Sec. 2 describes the method. We first discuss the appearance of the target brain structures in Sec. 2.1. This motivates the selection of candidate regions described in Sec. 2.2. We then define a region description model in Sec. 2.3. Section 2.4 describes the machine learning-based framework for region-classification (segmentation), with implementation details in Sec. 3. Section 4 presents the datasets used in the experiments. Validation experiments and

results are reported in Secs. 5 and 6, respectively. We conclude, in Sec. 7, with a discussion of the contributions of the work and highlight areas of possible future work.

## 2 Method

### 2.1 Target Fetal Brain Structures

The interaction of the US beam with a fetal brain structure results in an acoustic pattern response that either yields a bright anatomical mapping of the structure or alternatively a dark acoustic signature. The brightness of the signature depends on its echogenicity that an experienced sonographer learns to interpret proficiently. To automate structure detection and segmentation, we are interested in learning how to detect, segment, and characterize “blob-like” acoustic responses that correspond to structures of interest (SOIs). Specifically, in this article, we focus on two brain structures, the CP and the CC. As shown in Fig. 2, the CP and CC differ in terms of intensity, shape, and configuration (spatial arrangement) and hence provide two excellent examples on which to test the automatic detection and characterization framework.

### 2.2 Candidate Region Selection

We assume that, in a given US image, a fetal brain structure appears as a bright or dark region depending on the echogenicity of its tissue. Tissue with low echogenicity presents as a dark blob, and vice versa. Both CC and CP are examples of structures that are approximately homogeneous, which leads to a first assumption of our framework, namely that a target of interest is a local extremal region regardless of its echogenicity (whether it is a local maximum or minimum). Furthermore, it is intuitive to assume that the ideal segmentation of such a fetal brain structure should be contiguous and stable over thresholding. Based on the second assumption, we model candidate regions as maximally stable extremal regions (MSERs).<sup>20</sup> Specifically, we define image  $I$  to be a mapping from the coordinate domain  $D$  to the intensity domain. Region  $R$  is a contiguous subset of  $D$ , and  $\partial R$  is the region boundary. A pixel  $p$  belongs to  $R$ :  $p \in R$ , while  $q$  is a pixel that belongs to  $\partial R$ . An extremal region is either a maximal extremal region or minimal extremal region. We define  $R$  as a maximal extremal region if  $I(p) > I(q)$  for all  $p \in R$  and  $q \in \partial R$  [or  $I(p) < I(q)$  for minimal extremal region]. An important property of extremal regions is their nestedness, that is, if a set of extremal regions is sorted in a monotonic order of their intensity values

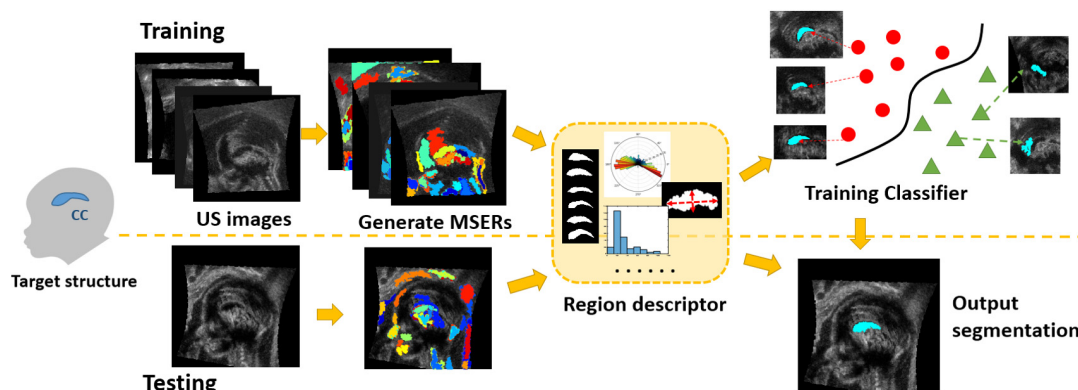
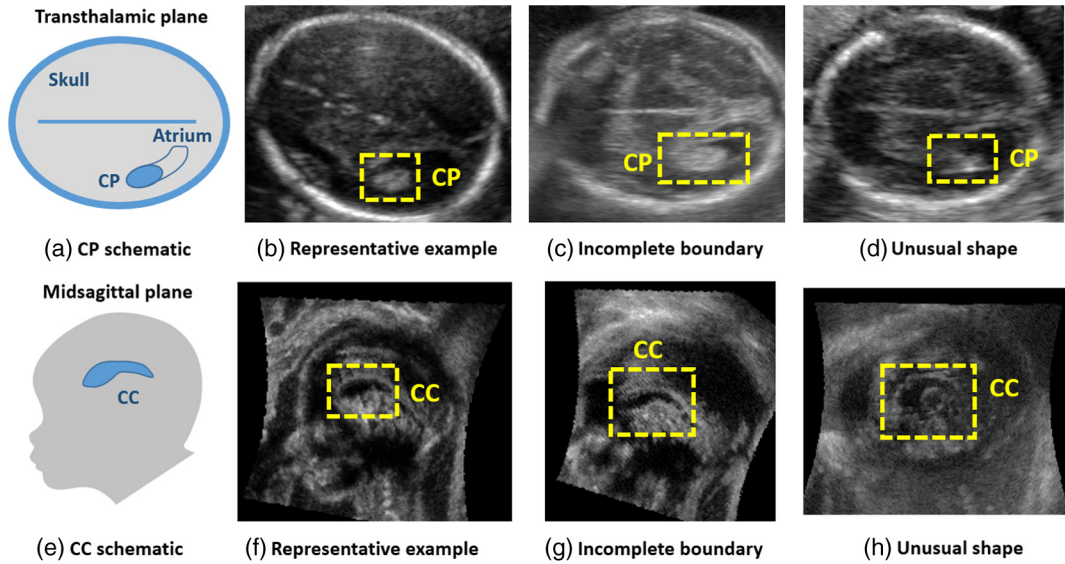


Fig. 1 Schematic flowchart of automated segmentation framework.



**Fig. 2** Transthalamic plane (first row) and midsagittal plane (second row) of the fetal brain. Schematics for CP and CC are shown in (a) and (e), respectively. Other subfigures in each row are corresponding US image examples. The CP or CC is bounded by a yellow dashed box in each image.

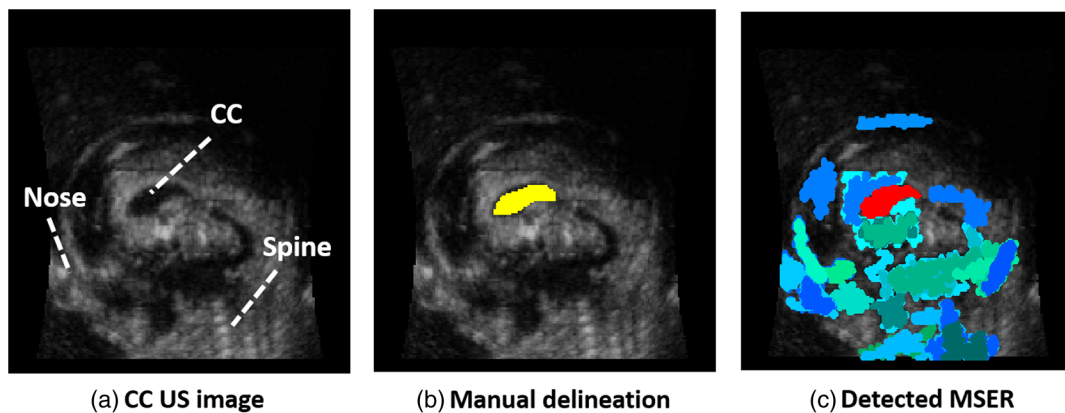
$R_1, \dots, R_i, R_{i+1}, \dots$ , then it follows that  $R_i \subset R_{i+1}$  for all  $R_i$ . Then, we choose the maximally stable region from the whole set of nested extremal ones. We define a function of  $i$

$$f(i) = (|R_{i+\delta}| - |R_{i-\delta}|) / |R_i|, \quad (1)$$

which represents the area variation between two nested regions, where  $\delta$  denotes their intensity interval and  $|\cdot|$  denotes area. It is intuitive that the region is more stable if its intensity changes aggressively over its boundary, which leads to a smaller value of  $f(i)$ . Conversely, a larger  $f(i)$  indicates instability as the area of nested extremal regions varies. Finally, an extremal region  $R_j$  is defined to be maximally stable (an MSER) if  $f(j)$  is a local minimum of  $f(i)$  and smaller than  $\gamma$ —a threshold that controls the stability of the selected region.

In practice, the aforementioned parameters for generating MSERs (controlling threshold  $\gamma$  and intensity interval  $\delta$ ) are chosen empirically and different problems usually have different

optimal parameter sets. For example, the CC may be best selected by parameter set  $S_a = \{\gamma_a, \delta_a\}$ , whereas the CP is best selected using a different parameter set  $S_b = \{\gamma_b, \delta_b\}$ . Tuning parameters for each specific task limits the applicability of the algorithm on different tasks, and is time-consuming. To improve the adaptability as well as to reduce human intervention, we generate multiple sets of MSERs using different values of  $\gamma$  and  $\delta$  sampled from a broad range of values. The machine learning algorithm selects the appropriate region from these sets of MSERs as described in Sec. 2.4. As a result, no human intervention is needed to select different parameters to segment different structures. Thus, the proposed framework readily adapts to a segmentation task. It is important to note that the MSER can be replaced with other region extraction methods to use our pipeline in segmenting objects from other imaging modalities. The MSER is chosen for this work for its great performance on the targeted fetal neurosonography. We also note that there may be one or more MSERs corresponding to the desired region (i.e.,



**Fig. 3** Candidate region detection. (a) The original fetal US image of midsagittal plane. The CC, the spine, and the nose of the fetus are pointed out. (b) The manual delineation of the CC in yellow. (c) The detected MSERs, where the positive example is plotted in red and the negative examples are randomly assigned with blue or green for comparison. It can be seen that the majority of MSERs belong to negative class.



true-positive examples) for each image as the MSERs were generated using different parameters.

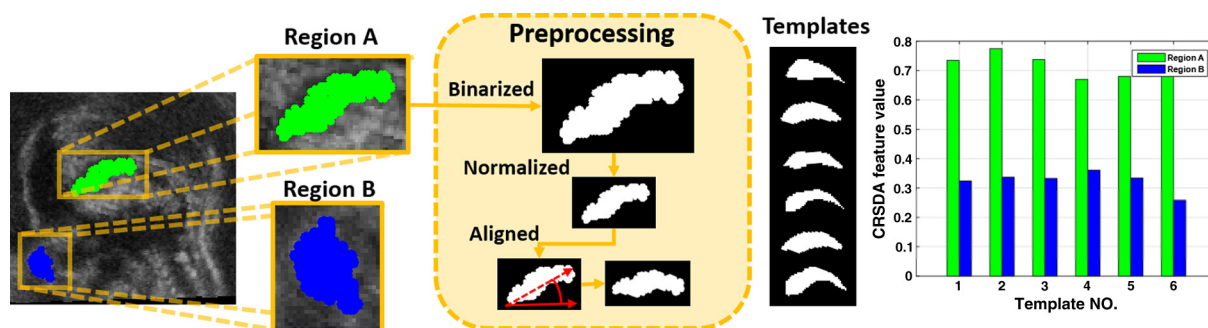
Detected MSERs (as illustrated in Fig. 3(c)) serve as candidate regions of the SOI and their features are extracted to form a region descriptor. As shown in Fig. 3(c), many MSER candidate regions are identified but within the pool of candidates lies a segmentation of the correct structure [shown in red in Fig. 3(c)]. To identify the correct structure, we construct a rich region-based feature descriptor. This is subsequently used within a machine-learning framework to identify the correct structure. We describe the region descriptor next.

### 2.3 Region Descriptor

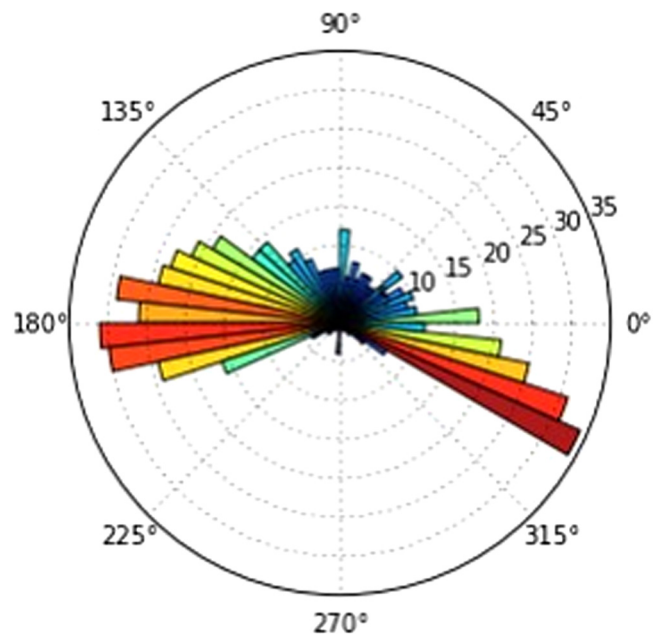
The shape of an anatomical brain structure is one of its most distinguishing features. It can be challenging to characterize fetal brain anatomy, where shape is usually irregular and has large intraclass variation. We characterize the shape of the SOI without an explicit model. All candidate regions are first preprocessed to remove irrelevant information (Fig. 4). To be more precise, the regions are binarized, resized, and aligned thus eliminating the effect of intensity, size, and orientation. The region width is scaled to the same size, whereas the aspect ratio is maintained to prevent distortion. Region alignment is achieved by rotation along the major axis to align their orientations with a common image axis. The shape of each candidate region is then described in two ways.

**Shape descriptor 1:** a set of shape templates is defined as a randomly selected subset of manual delineations. The delineations are preprocessed as described earlier.

The normalized crosscorrelation (NCC) between a preprocessed candidate region and each template is computed, with the maximum NCC value selected as a measure of shape similarity. Figure 4 shows the template-matching method applied to two separate regions. Region A (green) is the MSER generated from the desired SOI (i.e., CC, in this example), region B (blue) is a randomly chosen MSER belonging to the background (referring to Fig. 4). The column of templates shows six CC-shaped templates (randomly chosen for demonstration). The rightmost histogram compares the values of shape features extracted from regions A and B, respectively. The features of region A (green bars) have greater values than those of region B (blue bars) as region A's shape is more similar to the templates.



**Fig. 4** Shape descriptor 1: preprocessing procedure and feature extraction. Shape descriptor 1 is derived from NCC values, and its value is proportional to the shape similarity between candidate region and templates. The column of templates shows six examples of CC templates used in the experiments.



**Fig. 5** Shape descriptor 2: the area distribution of the candidate region on a polar coordinate system. The color red represents a large number of pixels lies within the fan section, while blue represents a small number.

**Shape descriptor 2:** A polar-histogram is constructed to describe the region area distribution centered on the region centroid. The polar region is divided evenly into 60 bins. Each bin contains a count of bright pixels (pixels within the region mask) that lie within the corresponding fan section of the polar coordinate system (refer to Fig. 5). This descriptor is motivated by Ref. 21 for cell detection, which used a method of boundary distribution instead. This descriptor provides 60 features.

Additionally, six other feature descriptors are derived from intensity, scale, and location information, namely

**Mean intensity.** It coarsely characterizes the intensity level of the region.

**Histogram of pixel intensities in the region.** The intensity distribution illustrated by this histogram characterizes both the region stability (intensity variation) and texture (repeated intensity pattern). It also complements

the region brightness information apart from the mean intensity as it is biased if large intensity variations exist. This descriptor generates 20 features.

**Length of the minor and major axes of the region.**

These approximate the region dimensions while being rotation-invariant. The minor and major axes are derived from the smallest ellipse that enclosed the region.

**Eccentricity of the region.** It is the ratio of the length of the minor axis to the major axis. A smaller eccentricity indicates a more circular region, whereas a larger value indicates an elongated shape.

**X and Y coordinates of the region centroid.** These denote the location of the candidate within the image.

**Histograms of the normalized X and Y components of the region pixel co-ordinates.** Specifically, pixel coordinates within the region ( $x$ ,  $y$ -coordinates separately) divided by the size of image. This descriptor provides 60 features.

In total, this gives a region feature descriptor with 190 terms.

## 2.4 Classification and Voting Mechanism

To preserve the generality of the algorithm, the image is not pre-cropped and minimal human intervention is allowed during parameter selection (explained in Sec. 2.2). As a result, candidate region selection generates a larger proportion of negative class responses relative to positive ones, e.g., 28 negative versus 1 positive case in Fig. 3(c). This results in an imbalanced dataset, which may bias the classifier. To accommodate this, we employ Random Under-Sampling (RUSBoost)—an iterative boosting algorithm designed to deal with data imbalance for classification.<sup>22</sup> In this method, random under-sampling is applied to remove the majority class examples until the desired class ratio is achieved. Specifically, at iteration  $t$ , a temporary training dataset  $S_t$  is created from the dataset  $S$  using random under-sampling to remove majority class examples until a desirable percentage (e.g., 50%) of the minority class is achieved. For each candidate region  $R_i$ , the region descriptor extracts its features and maps it as a feature point  $x_i$  in feature space  $X$ . Let  $y_i \in Y$  be the class label of  $x_i$  and  $D_t(x_i)$  be the weight of  $x_i$  in iteration  $t$ . A simple decision tree (weak-learner)  $G_t(x_i): X \rightarrow Y \in [0,1]$  is then trained based on  $S_t$ . The loss function  $e_t$  is defined as

$$e_t = \sum_{i=1}^n D_t(x_i) \mathbb{1}[G_t(x_i) \neq y_i], \quad (2)$$

with  $\mathbb{1}(\cdot)$  is an indicator function.

In the next step, the weight  $D_t(x_i)$  is updated as

$$D_{t+1}(x_i) = \frac{D_t(x_i)}{Z_t} \alpha_t [y_i G_t(x_i)], \quad (3)$$

where  $Z_t$  is the normalizing factor and  $\alpha_t$  is given by

$$\alpha_t = \frac{1}{2} \log \frac{1 - e_t}{e_t}. \quad (4)$$

This process increases the weights of misclassified examples and decreases those of correctly classified ones. After

training is complete, all weak learners are combined to yield the final classifier. On rare occasions, US artifacts may possess similar features (e.g., intensity or shape) to the SOI, which can lead to a confusing prediction (i.e., false-positive examples). However, our experiments have shown that those artifacts are much less stable across different values of the parameters.

To combine coexisting positive examples and eliminate artifacts, a simple voting mechanism is applied after region classification. Specifically, every candidate that is classified as positive casts a vote on the corresponding region. The region with maximal votes is selected as the SOI after accumulating all the votes.

## 3 Implementation Details

The stability threshold and intensity interval were set as  $\gamma = 0.1$  to 0.4 (step size = 0.05),  $\delta = 0$  to 5 (step size = 0.5). The number of RUSBoost iterations was set as  $t = 200$  and a learning rate of 0.25 was chosen empirically. The average runtime for the segmentation for one CP image is 12.1 s, while for a CC image is 12.6 s on a 3.5 GHZ, 16 GB RAM computer.

## 4 Datasets

Experiments were conducted on US datasets of two fetal brain structures of clinical importance, CP and CC, respectively. The CP was selected as it plays a primary role in the formation of cerebrospinal fluid and is associated with several congenital diseases.<sup>23,24</sup> The fetal CC is considered as a sensitive biomarker of normal brain development and maturation.<sup>25</sup> Indeed, malformation and agenesis of the CC have been frequently related to several genetic diseases.<sup>26,27</sup> The image datasets that contained each of these structures are introduced next.

**CP Dataset:** The CP appears as a bright circular region in the trans-thalamic (TT) US image plane of the fetal brain, with varying shape and size across gestation (Fig. 2). The dataset we have used consists of 120 2-D US images of the standard TT plane of the fetal brain, which were randomly selected from a large clinical study database of healthy subjects.<sup>28</sup> The images were of healthy fetuses between 20 and 22 gestational weeks corresponding to the period of clinical anomaly scanning for Trisomy 18.<sup>29</sup> The images were acquired using a 2-D linear probe (Phillips HD9) at 2- to 5-MHz wave frequency and at a pixel resolution of 2 mm  $\times$  2 mm. Data were divided into a training group of 68 images and a testing group of 52 images.

**CC Dataset:** The CC is observed as a single comma-shaped echogenic structure on the midsagittal plane of fetal US brain image (Fig. 2). Our dataset consisted of 219 2-D US images of the midsagittal plane of the fetal brain. The subjects were randomly selected from 21- to 30-gestational week fetuses from the same database as described above.<sup>28</sup> The typical size of each image was 210  $\times$  170 pixels, with each isotropic pixel measuring 0.6 mm in size. Images were divided into a training group (137 images) and a testing group (82 images).

## 5 Experiments

### 5.1 Classification and Feature Comparison

We evaluated the performance of three different classifiers for CC and CP segmentation, respectively. Specifically, we compared RUSBoost with support vector machines (SVMs) and AdaBoost. The systematic error of different classifiers is reported in terms of accuracy, sensitivity, specificity, and precision (Table 1).

To demonstrate the effectiveness of the region descriptor, RUSBoost classification experiments were conducted using different feature sets such as shape features only, the appearance and location features (referred as “other features” in Fig. 8), and the combination of both. Size and shape of a brain structure change during brain development. We investigated whether performance varied with gestational age (GA). The region classification accuracy was binned and compared based on the GA of the subjects.

**Table 1** Table of the performances assessed on the different classifiers. All indices are reported in percentage. The RUSBoost achieved that the best results for both CC and CP classification are shown in bold.

Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
SVM-CC	98.6	82.4	99.3	90.4
AdaBoost-CC	98.7	89.6	99.6	91.1
RUSBoost-CC	<b>99.4</b>	<b>97.2</b>	<b>99.6</b>	<b>94.5</b>
SVM-CP	96.6	86.6	99.6	92.4
AdaBoost-CP	97.9	89.8	99.5	95.2
RUSBoost-CP	<b>98.0</b>	<b>94.7</b>	<b>98.6</b>	<b>97.6</b>

**Table 2** Automatic segmentation accuracy compared with human experts’ results. Biometric measurements for CC (length) and CP (diameter) are reported in the first three rows. The measurement for CC has more significant digits as the original image resolution is higher than that of CP (CC: 0.6 mm, CP: 2 mm). Intra represents intraobserver error, whereas Inter stands for interobserver error. Auto is the comparison between the automated result and ground truth. The Hausdorff distance and DSC between two segments (automated versus manual) are displayed in the fourth and the fifth rows, respectively.

Brain	CC		CP	
	Median	Mean $\pm$ SD	Median	Mean $\pm$ SD
Structure				
Intra/(mm)	1.88	2.54 $\pm$ 2.23	5.6	5.9 $\pm$ 4.4
Inter/(mm)	2.62	2.97 $\pm$ 2.36	6.9	7.2 $\pm$ 5.2
Manual versus auto/(mm)	1.59	1.81 $\pm$ 1.40	5.9	6.7 $\pm$ 4.5
$d_H$ /(mm)	2.39	2.80 $\pm$ 1.91	7.3	7.7 $\pm$ 3.4
DSC	0.84	0.81 $\pm$ 0.06	0.79	0.76 $\pm$ 0.08

### 5.2 Segmentation Accuracy

We present both qualitative and quantitative evaluation experiments on segmentation accuracy. Figures 9 and 10 show typical CP and CC segmentation results.

Table 2 shows automatic segmentation accuracy using two statistics commonly used to compare the similarity of objects: dice similarity coefficient (DSC)<sup>30</sup> and Hausdorff distance ( $d_H$ ).<sup>31</sup>

Automatically derived linear measurements were compared with manual measurement by two human observers. Following International Society of Ultrasound in Obstetrics and Gynecology guidelines, the length of the CC and the diameter of CP were both manually measured by two different observers (to estimate interobserver variability) and one observer repeated manual measurement to estimate intraobserver variability. The third column of Fig. 10 shows five examples of CC length measurement. Figure 11 and Table 2 compare the inter/intraobserver variability and the deviation of automated results from the ground truth.

## 6 Results

### 6.1 Classification Performance

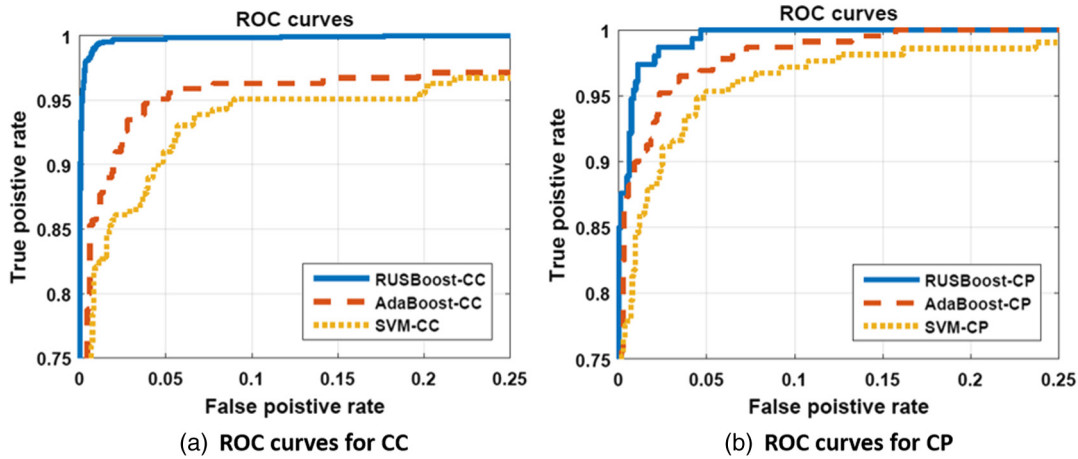
Table 1 shows that the algorithm achieves good accuracy for both CC (Acc = 99.4%, Prec = 94.5%) and CP (Acc = 98.0%, Prec = 97.6%) in region identification. All the models had high specificity (= true negative rate). The value of sensitivity (= true positive rate) is interesting as a high sensitivity indicates that the algorithm is not over-fitted to the dominant class and can generalize well on unseen data. The RUSBoost model achieved the best sensitivity compared with AdaBoost and SVM (CC: 97.2%, CP: 94.7%).

The receiver operating characteristic (ROC) curves for each classifier are plotted in Fig. 6. It shows that the RUSBoost classifier achieved the highest area under the curve (AUC), whereas AdaBoost performed slightly better than SVM.

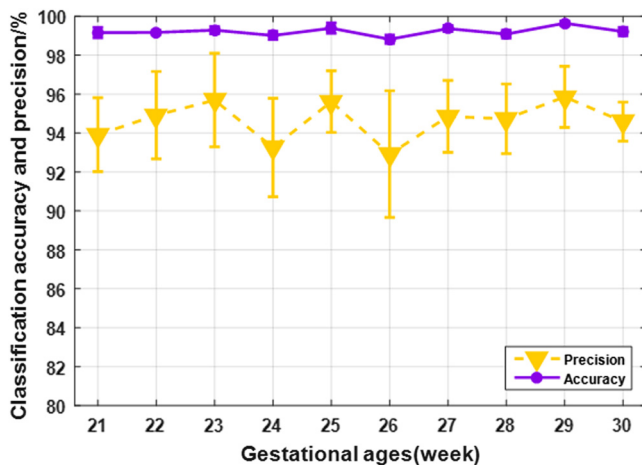
As the anatomical structures grow rapidly in the developing brain, their appearance and image contrast can vary greatly across gestation. We also investigated the influence of GA on the performance of classifiers using the CC dataset as it had larger variation in terms of GA (from 21 to 33 weeks). Figure 7 shows classification accuracy and precision with respect to weeks of gestation for CC classification. It can be seen that the variation of classifier performance is <2% in accuracy and <6% in precision. This shows our framework is robust to typical structure appearance and image contrast variation due to GA.

### 6.2 Feature Set Comparison

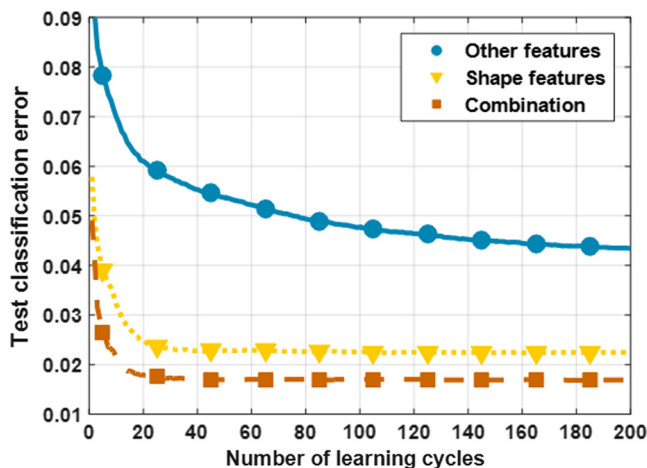
Figure 8 shows classification error using different feature sets. It shows that classification based on shape features alone achieved a lower error rate,  $e = 2.2\%$  at the  $t = 200$  iteration than that of the appearance and location features,  $e = 4.3\%$ . The combination achieved the lowest error rate,  $e = 1.9\%$ . Additionally, it can be seen (Fig. 8) that the algorithm converged quickly with shape features alone (around the 30<sup>th</sup> iteration), whereas the algorithm was not fully converged at iteration 200 using only the appearance and location features. The fact that the classification accuracy relies heavily on the shape feature, supports our hypothesis in Sec. 2.3 that the shape is one of the most distinguishing characteristics in fetal brain (structures) segmentation.



**Fig. 6** Receiver operating characteristic (ROC) curves for different classifiers, namely SVM, Adaboost, and RUSBoost assigned with different SOI: (a) CC and (b) CP segmentation tasks. The blue solid line represents RUSBoost classifier, whereas the curve for AdaBoost is plotted using orange dashed line and yellow dotted line for SVM.



**Fig. 7** The region classification accuracy (purple) and precision (yellow) are plotted with respect to fetal GA. The number of volumes ranged from 15 to 25 per GA week.



**Fig. 8** The classification error using different feature sets is plotted against learning iterations (shape features: yellow, other feature: blue, and the combination: orange).

### 6.3 Segmentation Accuracy

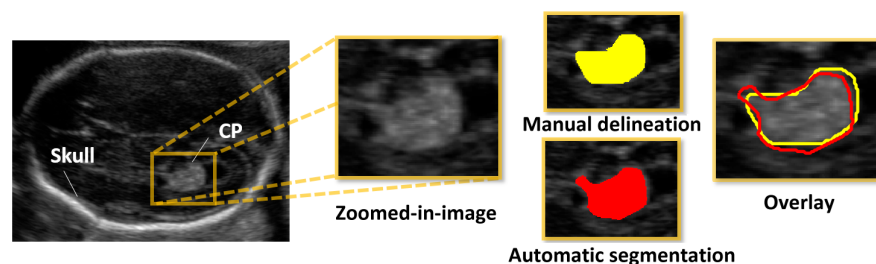
Previous section shows our algorithms were able to segment SOIs given different imaging conditions. In addition, automatic delineation corresponded well to the manual delineation. Figure 9 shows a typical CP segmentation result (automatic: red, manual: yellow) with CP overlaid on the TT US image. Note that the automatic delineation segments the CP well from the background. Note further that in the upper-left part of the CP, where the two delineations disagree, the human expert made a subjective judgment about boundary completeness.

Figure 10 shows CC segmentation results for five different cases; the first column shows the original US images, and the second column shows the corresponding automatic segmentations in light blue. Cases 1 and 2 in Fig. 10 show how the approach is able to deal with large intraclass variation of the shape of the SOI. The framework is also able to accommodate varying contrast (cases 3 and 4), acoustic shadows (the head of CC in case 4), and a low signal-to-noise ratio (case 5).

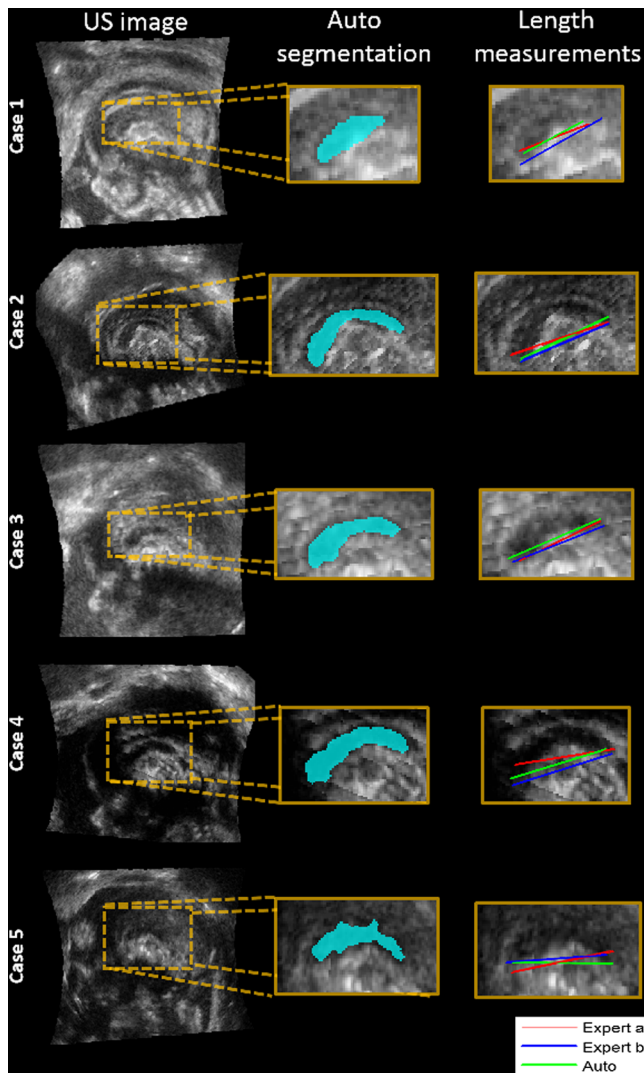
The third column of Fig. 10 compares biometry measurement carried out by two different human experts and those derived from automatic segmentation. The three lines (expert A: red line, expert B: blue line, automatic result: green line) coincide well with each other for all cases. The largest disagreement occurred in case 4, where the head of the CC is occluded by acoustic shadows. In this case, the human observer made subjective assumptions about the CC boundary. By contrast, our algorithm is able to capture small intensity variation, which may be imperceptible by eye and offers a more stable measurement.

Figure 11 shows boxplots of inter/intraobserver variability and the deviation of automated results from the ground truth. We see that the automated versus manual errors, for both CC and CP, are comparable with those of the intra/interobserver error margins. Table 2 summarizes the results numerically. It can be seen that the automated versus manual error for CC length measurements is  $1.81 \pm 1.40$  mm, which is slightly smaller than the intraobserver deviation  $2.54 \pm 2.23$  mm. The average error value for automated CP diameter measurement lies within intra and interobserver deviation. This further shows that the designed framework can accommodate large intraclass variations in intensity, size, and location of one





**Fig. 9** Manual and automatic segmentation of CP. Original US image is shown in the far-left. Subimage of CP is enlarged to visualize the details. Manual delineation (yellow) and automatic generated segmentation (red) of CP are also displayed. The two delineations are plotted in the right-most image for comparison.



**Fig. 10** CC automatic segmentation results for five different cases. Each row belongs to one patient. The first column are the original US images, the corresponding automatic segmentations are presented in the second column in light blue. The third column shows the CC length measurements carried out by two different human experts and those derived from automatic segmentation (expert a: red line, expert b: blue line, automatic result: green line).

type of brain structure and is able to generate measurements that are similar to manually derived ones. The segmentation accuracy is also evaluated quantitatively as shown in Table 2. DSC values of CC segmentation  $0.81 \pm 0.06$  is slightly higher

than that of the CP  $0.76 \pm 0.08$ . This may be explained by the highly varying spatial appearance of the CP [Figs. 2(b)–2(d)]. The values of  $d_H$  ( $2.80 \pm 1.91$  mm CC,  $7.7 \pm 3.4$  mm CP) are competitive with respect to interobserver error of linear measurements ( $2.97 \pm 2.36$  mm CC,  $7.2 \pm 5.2$  mm CP). It proves that the automatically generated segmentations correlate well with the manual delineations.

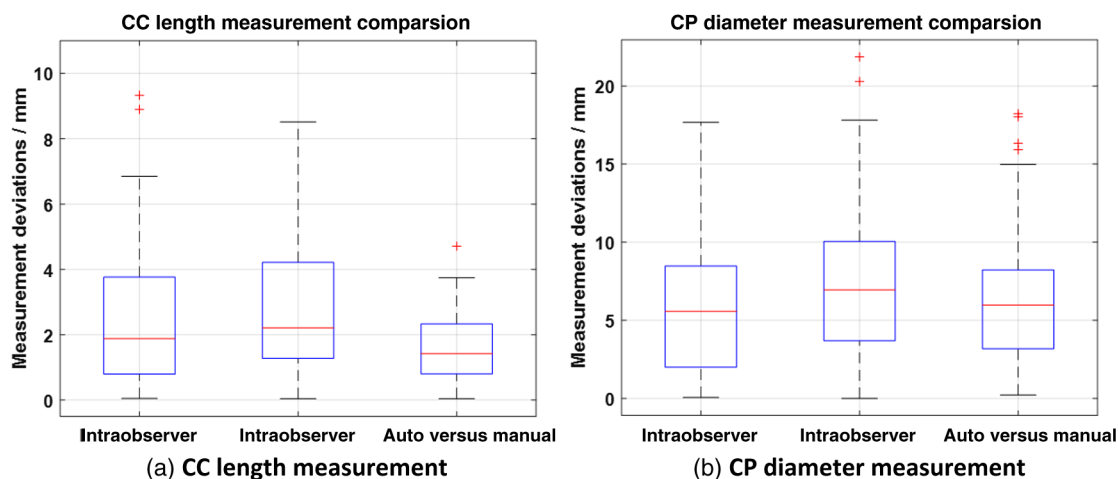
Furthermore, recall that the parameters used in the framework are not manually tuned separately for segmenting the CP or CC (the two experiments used the same parameter sets as given in Sec. 3), yet it performed well for both tasks. This suggests that the method can adapt to differences in appearance, shape, and spatial configuration of an SOI.

## 7 Discussion

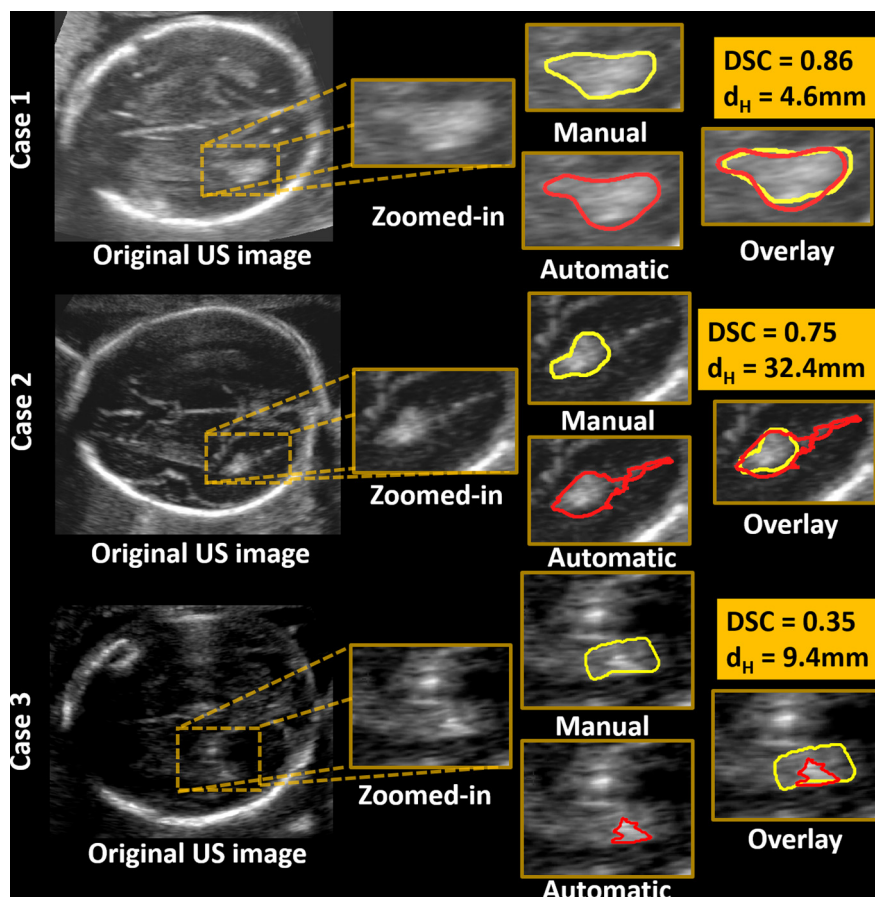
We have reported experiments and their results to evaluate the proposed framework. In this section, we analyze the rationale behind our experimental results and discuss two interesting failure cases.

We compared the performance of different classifiers in CC and CP task, respectively, and concluded that the RUSBoost achieved the best performance (Sec. 6.1). We discovered that the difference in performance between the three classifiers is greater for CC classification than that for CP (Fig. 6). Our interpretation of this is that CC classification is a more imbalanced classification problem: only 7.6% positive examples among the whole dataset, whereas the CP has 15.44% positive examples. It results from a more complex anatomical pattern in the midsagittal plane than that of the transthalamic plane (i.e., more MSERs belong to the negative class). This shows that the RUSBoost classifier can handle the class imbalance problem superiorly and is preferable to other common classifiers for our problem.

The shape descriptor demonstrated its importance in assisting region classification (Sec. 6.2). Nonetheless, intensity and location features also provide additional information to assist classification. The first example in Fig. 12 shows a challenging case of CP segmentation. Although the CP in that image has an unusual shape, the candidate region is still classified as positive as its intensity and location features have high prediction confidence. The resulting automatic segmentation matches well with the manual result. Case 2 in Fig. 10 is similar as its CC has an unusual elongated shape, yet the algorithm segments the SOI and autobio metric measurements coincides well with that of the expert. It might be useful to explore other features, suggested by these two examples, as the imaging condition and structure appearance can change



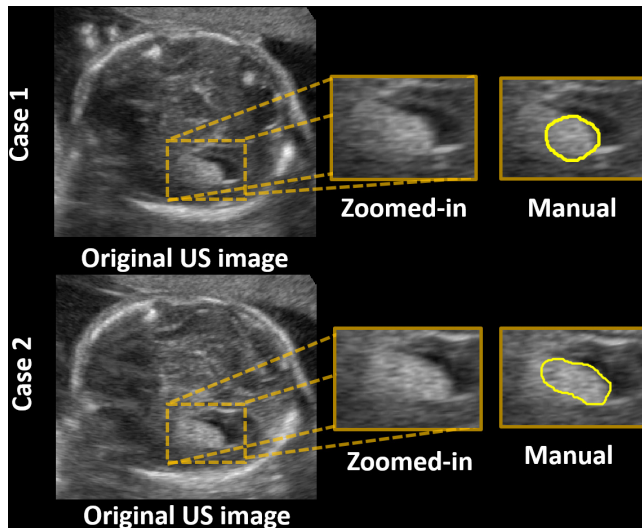
**Fig. 11** Biometric measurements evaluation. (a) The boxplot of measurements deviation of CC length between observers and automated results and (b) counterpart boxplot for CP diameter measuring.



**Fig. 12** CP automatic segmentation results. Each row displays a single subject. The CP region and the segmentation results are enlarged for easier examination. The manual segmentation is plotted in yellow, whereas the automatic result is plotted in red. The two outlines (for each case) are then overlay on the same image for comparison. The value of the DSC and  $d_H$  is also displayed in the upper-right corner.

greatly in fetal neurosonography. Moreover, there is some interesting literature on selecting/merging the most relevant features or comparing their performance.<sup>32-35</sup> Future works can explore this area in more detail to further enhance the performance.

The process of designing the region descriptors requires understanding and prior knowledge of US image and fetal brain anatomy. Recent success of neural networks (NNs) for large-scale computer vision tasks<sup>36-38</sup> has led to a shift away from handcrafted features. Although NNs report excellent



**Fig. 13** The two failure case of CP segmentation task within the test dataset. The CP image is enlarged for closer examination. The manual delineation is plotted in yellow.

performance on large annotated datasets,<sup>39–42</sup> the proposed region descriptors are better suited and a simpler way for solving the presented task.

### 7.1 Challenging Cases

To demonstrate the algorithm performance on challenging cases, we further discuss the cases with the largest Hausdorff distance and the smallest DSC within the entire sets (refer to cases 2 and 3 in Fig. 12). As the CP is connected with a thin bright structure in the original US image in case 2, the automatic segmentation looks almost like a shooting star. As its tail section only appears in a narrow region in the upper-right direction, only one or two feature values would be different from that produced by the manual delineation using the described region descriptor. The result recorded high Hausdorff distance while still reporting reasonable DSC since its head correlated well with the manual delineation. It should be noted that the human observer ignored the “tail.” Segmentation methods that use predefined models may avoid this situation, but their adaptability toward varying anatomy shape is weak. Our method’s robustness toward connected small structures might be improved by adding a postprocessing module in the future.

Case 3 in Fig. 12 has the smallest DSC value as the algorithm picks up a small bright region within the manual segmentation. This behavior is rational as the CP nearly merges with the surrounding bright tissue in the original image, whereas the small triangular region in the bottom right of the image satisfies the intensity and location criteria of a typical CP. In addition, this result demonstrates that our method does not have strict preselection criteria in candidate region selection.

### 7.2 Failure Cases Analysis

We report on two interesting cases of CP segmentation in Fig. 13, where our framework failed. It can be seen that the CP and surrounding tissues form a homogeneous bright region in both cases. Thus, the algorithm is not able to detect a suitable candidate region. We note that the enlarged CP area (according to the manual delineations) of the two cases has a substantially

similar appearance, yet the same human observer delineates them quite differently (a circular outline versus a elliptical outline). The reason for the disagreement between the two manual delineations can be explained by the relatively poor contrast of the SOI with respect to the background and nearly half of the boundary delineation relied on subjective assumptions. This observation, however, echoes the previous view that manual segmentation for this task is both hard and subjective.

## 8 Summary

This paper has presented a general method for automatic brain structure segmentation in fetal neurosonography. In particular, the shape descriptor is designed to be invariant to intensity, scale, and rotation and to capture shape characteristics. The resulting segmentations were found to match manual delineations with high accuracy. Experiments on CC and CP segmentation require a different setting (midsagittal versus transthalamic plane) of the developing brain. The results demonstrate that our method is generalizable to different brain structures without the need to develop a tool for each of them individually. The method performed well on clinical data over a broad GA range, and is fully automatic. Future research will focus on morphometric analysis of fetal brain structures using the derived segmentations and the application of study on fetal neurological development.

### Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

### Acknowledgments

We acknowledge the use of fetal neurosonography datasets from the UK site of *Intergrowth-21st study* for this research. This work was supported by the EPSRC Programme Grant Seebibyte (EP/M013774/1) and the National Institutes of Health (NIH) through National Institute on Alcohol Abuse and Alcoholism (NIAAA) (2 U01 AA014809-14). Dr. Namburete was supported by a Royal Academy of Engineering Research Fellowship.

### References

1. S. Carroll et al., “Correlation of prenatal ultrasound diagnosis and pathologic findings in fetal brain abnormalities,” *Ultrasound Obstet. Gynecol.* **16**(2), 149–153 (2000).
2. L. R. Pistorius et al., “Fetal neuroimaging: ultrasound, MRI, or both?” *Obstet. Gynecol. Surv.* **63**(11), 733–745 (2008).
3. ISUOG, “Sonographic examination of the fetal central nervous system: guidelines for performing the basic examination and the fetal neurosonogram,” *Ultrasound Obstet. Gynecol.* **29**(1), 109–116 (2007).
4. P. Hilpert, B. Hall, and A. Kurtz, “The atria of the fetal lateral ventricles: a sonographic study of normal atrial size and choroid plexus volume,” *Am. J. Roentgenol.* **164**(3), 731–734 (1995).
5. E. A. Reece et al., “Fetal cerebellar growth unaffected by intrauterine growth retardation: a new parameter for prenatal diagnosis,” *Am. J. Obstet. Gynecol.* **157**(3), 632–638 (1987).
6. G. Carneiro et al., “Semantic-based indexing of fetal anatomies from 3-D ultrasound data using global/semi-local context and sequential sampling,” in *IEEE Conf. on Computer Vision and Pattern Recognition* (2008).
7. A. Namburete, B. Rahmatullah, and J. A. Noble, “Nakagami-based choroid plexus detection in fetal ultrasound images using adaboost,” in *Proc. of the Medical Image Understanding and Analysis*, pp. 31–36 (2012).



8. M. Sofka et al., "Automatic detection and measurement of structures in fetal head ultrasound volumes using sequential estimation and integrated detection network (IDN)," *IEEE Trans. Med. Imaging* **33**(5), 1054–1070 (2014).
9. I. Zador et al., "Ultrasound measurement of the fetal head: computer versus operator," *Ultrasound Obstet. Gynecol.* **1**(3), 208–211 (1991).
10. G. K. Matsopoulos and S. Marshall, "Use of morphological image processing techniques for the measurement of a fetal head from ultrasound images," *Pattern Recogn.* **27**(10), 1317–1324 (1994).
11. W. Lu, J. Tan, and R. Floyd, "Automated fetal head detection and measurement in ultrasound images by iterative randomized Hough transform," *Ultrasound Med. Biol.* **31**(7), 929–936 (2005).
12. M. Yaqub et al., "Volumetric segmentation of key fetal brain structures in 3D ultrasound," in *Int. Workshop on Machine Learning in Medical Imaging*, pp. 25–32, Springer (2013).
13. B. Gutiérrez-Becker et al., "Automatic segmentation of the fetal cerebellum on ultrasound volumes, using a 3d statistical shape model," *Med. Biol. Eng. Comput.* **51**(9), 1021–1030 (2013).
14. P. A. Habas et al., "Atlas-based segmentation of developing tissues in the human brain with quantitative validation in young fetuses," *Hum. Brain Mapp.* **31**(9), 1348–1358 (2010).
15. A. Gholipour et al., "Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly," *Neuroimage* **60**(3), 1819–1831 (2012).
16. R. Wright et al., "Age dependent fetal MR segmentation using manual and automated approaches," in *MICCAI Workshop on Perinatal and Paediatric Imaging*, pp. 97–104 (2012).
17. Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3d brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(10), 1744–1757 (2010).
18. S. Eskildsen et al., "Beast: brain extraction using multiresolution non-local segmentation," in *MICCAI Workshop on Multi-Atlas Labeling and Statistical Fusion*, pp. 97–108 (2011).
19. M. Liu et al., "Spatially adapted augmentation of age-specific atlas-based segmentation using patch-based priors," *Proc. SPIE* **9034**, 90341H (2014).
20. J. Matas et al., "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vision Comput.* **22**(10), 761–767 (2004).
21. C. Arteta et al., "Learning to detect cells using non-overlapping extremal regions," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, pp. 348–356, Springer (2012).
22. C. Seiffert et al., "Rusboost: a hybrid approach to alleviating class imbalance," *IEEE Trans. Syst. Man cybern. A* **40**(1), 185–197 (2010).
23. L. Chan et al., "A sonographic and karyotypic study of second-trimester fetal choroid plexus cysts," *Obstet. Gynecol.* **73**(5 Pt 1), 703–706 (1989).
24. S. Gabrielli et al., "The clinical significance of prenatally diagnosed choroid plexus cysts," *Am. J. Obstet. Gynecol.* **160**(5), 1207–1210 (1989).
25. P. Goodyear et al., "Outcome in prenatally diagnosed fetal agenesis of the corpus callosum," *Fetal Diagn. Ther.* **16**(3), 139–145 (2001).
26. F. L. Bookstein et al., "Corpus callosum shape and neuropsychological deficits in adult males with heavy fetal alcohol exposure," *Neuroimage* **15**(1), 233–251 (2002).
27. L. Sztriha, "Spectrum of corpus callosum agenesis," *Pediatr. Neurol.* **32**(2), 94–101 (2005).
28. A. T. Papageorgiou et al., "International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21st project," *Lancet* **384**(9946), 869–879 (2014).
29. S. R. Turner et al., "Sonography of fetal choroid plexus cysts detection depends on cyst size and gestational age," *J. Ultrasound Med.* **22**(11), 1219–1227 (2003).
30. L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology* **26**(3), 297–302 (1945).
31. F. Hausdorff, *Mengenlehre*, Walter de Gruyter, Berlin (1927).
32. M. Unser and M. Eden, "Multiresolution feature extraction and selection for texture segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **11**(7), 717–728 (1989).
33. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(10), 1615–1630 (2005).
34. A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *IEEE 11th Int. Conf. on Computer Vision*, pp. 1–8, IEEE (2007).
35. T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Found. Trends Comput. Graphics Vision* **3**(3), 177–280 (2008).
36. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. of 25th Int. Conf. Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012).
37. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).
38. K. He et al., "Deep residual learning for image recognition," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).
39. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer (2015).
40. G. Litjens et al., "A survey on deep learning in medical image analysis," arXiv:1702.05747 (2017).
41. D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Ann. Rev. Biomed. Eng.* **19**, 221–248 (2017).
42. H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique," *IEEE Trans. Med. Imaging* **35**(5), 1153–1159 (2016).

**Ruobing Huang** is a PhD student at the Institute of Biomedical Engineering (IBME) at the University of Oxford, under the supervision of professor Alison Noble. Her research is focused on machine learning and its application on fetal neurosonography.

**Ana Namburete** is a Royal Academy of Engineering (RAEng) research fellow at the Department of Engineering Science at the University of Oxford and an associate research fellow at St. Hilda's College (Oxford). She completed a doctorate in IBME at the University of Oxford in 2011.

**Alison Noble** OBE FREng FRS is the Technikos professor of Biomedical Engineering at the Oxford University Department of Engineering Science, associate head of MPLS Division (Industry and Innovation), and a fellow of St. Hilda's College, Oxford. She is a former director of the Institute of Biomedical Engineering (2012–16). She is a fellow of the Royal Society, fellow of the IET, a fellow of the MICCAI Society, and a fellow of the Royal Academy of Engineering.