Automatic Detection and Measurement of Structures in Fetal Head Ultrasound Volumes Using Sequential Estimation and Integrated Detection Network (IDN)

Michal Sofka, *Member, IEEE,* Jingdan Zhang, Sara Good, S. Kevin Zhou, *Senior Member, IEEE,* and Dorin Comaniciu *Fellow, IEEE*

***Abstract*—Routine ultrasound exam in the second and third trimesters of pregnancy involves manually measuring fetal head and brain structures in 2D scans. The procedure requires a sonographer to find the standardized visualization planes with a probe and manually place measurement calipers on the structures of interest. The process is tedious, time consuming, and introduces user variability into the measurements. This paper proposes an Automatic Fetal Head and Brain (AFHB) system for automatically measuring anatomical structures from 3D ultrasound volumes. The system searches the 3D volume in a hierarchy of resolutions and by focusing on regions that are likely to be the measured anatomy. The output is a standardized visualization of the plane with correct orientation and centering as well as the biometric measurement of the anatomy. The system is based on a novel framework for detecting multiple structures in 3D volumes. Since a joint model is difficult to obtain in most practical situations, the structures are detected in a sequence, one-by-one. The detection relies on Sequential Estimation techniques, frequently applied to visual tracking. The interdependence of structure poses and strong prior information embedded in our domain yields faster and more accurate results than detecting the objects individually. The posterior distribution of the structure pose is approximated at each step by sequential Monte Carlo. The samples are propagated within the sequence across multiple structures and hierarchical levels. The probabilistic model helps solve many challenges present in the ultrasound images of the fetus such as speckle noise, signal drop-out, shadows caused by bones, and appearance variations caused by the differences in the fetus gestational age. This is possible by discriminative learning on an extensive database of scans comprising more than two thousand volumes and more than thirteen thousand annotations. The average difference between ground truth and automatic measurements is below 2 mm with a running time of 6.9 seconds (GPU) or 14.7 seconds (CPU). The accuracy of the AFHB system is within inter-user variability and the running time is fast, which meets the requirements for clinical use.**

***Index Terms*—object detection, sequential sampling, fetal head measurements, fetal brain measurements, fetal utrasound, 3D ultrasound.**

Michal Sofka is with Cisco Systems, Prague 2 12000 Czech Republic (msofka@cisco.com). This paper was completed while he was at Siemens Corporate Technology.

Jingdan Zhang is with Microsoft, Redmond, WA 98052 USA. This paper was completed while he was at Siemens Corporate Technology.

S. Kevin Zhou and Dorin Comaniciu are with Imaging and Computer Vision, Siemens Corporation, Corporate Technology, Princeton, NJ 08540 USA.

Sara Good is with Ultrasound Division, Siemens Healthcare, Mountain View, CA 94043 USA.

I. Introduction

Ultrasound is the most common imaging modality in obstetrics and gynecology [1] allowing real-time visualization and examination of the fetus. In the common practice of 2D exams, practitioners are trained to mentally reconstruct a 3D anatomy based on images of standard 2D planes resulting in high variations of biometric measurements depending on skill and training [2]–[4]. The most recent breakthrough in ultrasound imaging has come with the increasing use of systems capable of acquiring 3D volumetric data. The advantage of such data is in visualization of planes in which 2D acquisition is not possible, surface rendering of anatomies, and post-exam data processing and review [2]. The 3D acquisition decreases the examination time [4], [5] and reduces inter- and intra-observer variations of biometric measurements [3], especially for less experienced sonographers. Similar to 2D exams, the major bottleneck remains the navigation to the standardized planes [6], in which measurements are performed. For example, to find the ventricular plane in the fetal brain, a clinician needs to find cavum septi pellucidi, frontal horn, atrium, and choroids plexus (Figure 1). Finding the 2D planes in a 3D volume is tedious and time consuming. The process introduces a learning curve for an ultrasound specialist, who needs to recognize artifacts caused by a 3D reconstruction and understand the effects of various settings [2]. After the correct plane is found, the measurements are typically performed by manually placing calipers at specific landmarks of the measured anatomy.

This paper describes a system for Automatic Fetal Head and Brain (AFHB) measurements from 3D ultrasound. The standardized measurements can be used to estimate the gestational age of the fetus, predict the expected delivery date, assess the fetal size, and monitor growth. The input to the AFHB system is an ultrasound volume and a name of an anatomical part to be measured. Specifically, the parts we focus on here are fetal head and brain structures commonly measured and assessed in the second and third trimesters of pregnancy (see Figure 2 for examples). The output of the system is a visualization of the plane with correct orientation and centering as well as biometric measurement of the anatomy according to The International Society of Ultrasound in Obstetrics and Gynecology [6]. Such system will greatly reduce the time required to obtain fetal head and brain measurements, decrease the clinician strain, and minimize training associated with manual navigation to the standardized planes.

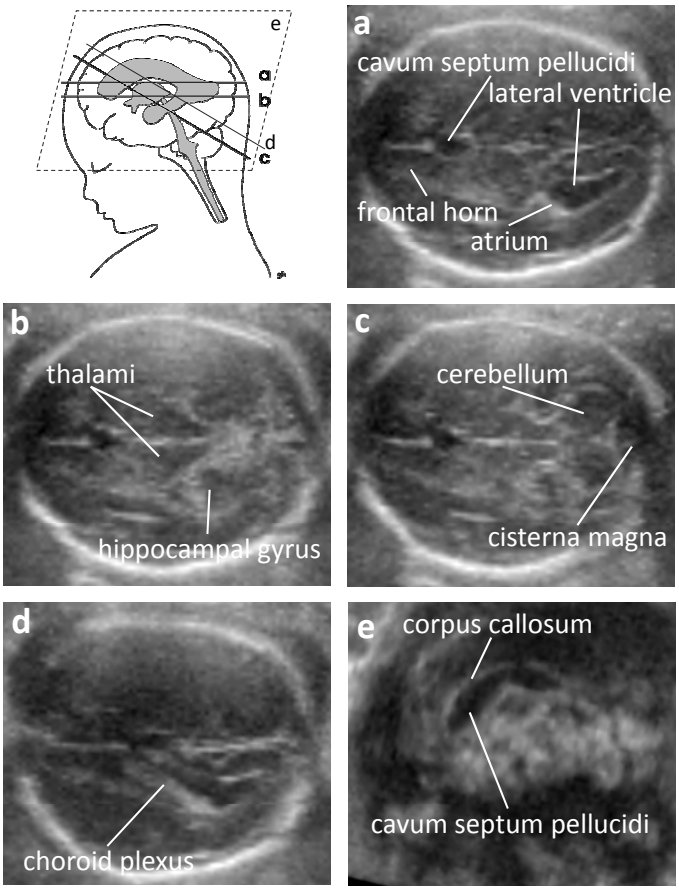

Fig. 1. Standard planes for performing the basic examination of the fetal nervous system and the fetal neurosonogram [6]: Ventricular plane (a), thalamic plane (b), cerebellar plane (c), choroid plexus plane (d), and median plane (e). Each plane is defined exactly by what anatomical landmarks and structures should be visible according to the guidelines of the International Society of Ultrasound in Obstetrics & Gynecology [6] (see Section II for details).

The state-of-the-art techniques for finding anatomical structures in 3D ultrasound images focus on automatic [7], [8] and semi-automatic [9] segmentation and detection [10], [11]. The most promising techniques are based on machine learning [10]–[13], pixel-wise classification [7], and deformable models [13], [14]. Although several existing detection algorithms can correctly identify structures of interest, they typically do not provide the visualization plane (standardized planes are especially important in obstetrics), do not compute the anatomical measurements, or they are too slow to be used in a clinical practice. Computer vision approaches for multi-object detection [15]–[17] rely on an individual detector for each object class followed by post-processing to prune spurious detections within and between classes. Individual object detectors can be connected by a spatial model to exploit relationships between objects [18]. Relative locations of the objects provide constraints that help make the system more robust by focusing the search in regions where the object is expected based on locations of the other objects. The most challenging aspect of these algorithms is designing detectors that are fast and robust, modeling the spatial relationships between objects, and determining the detection order. Structural learning approaches [19], [20], have been frequently applied to articulated 2D pose estimation [21]. Although structural dependency in the model can improve robustness, these approaches have not yet been applied to fast and accurate pose estimation in 3D images. The segmentation algorithms [7], [11]–[14] could, in principle, compute the measurements from the final segmentations, but they require the entire structure to be segmented which is slow and not as robust.

Although several successful techniques have been proposed for echocardiography [22], there are additional challenges in ObGyn applications. Similarly, the images have low signal-to-noise ratio, speckle noise, and other imaging artifacts. In addition, there are strong shadows produced by the skull and large intra-class variations because of differences in the fetus age. As gestational age increases, the skull bones develop, which makes it difficult for the acoustic signal to penetrate to the brain anatomy. This results in decreased detail in the image. Finally, some structures in the fetal brain are much smaller (less than 5 mm) than structures measured in echocardiography.

In this paper, we propose a detection system that addresses the challenges above. Our approach is motivated by Sequential Estimation techniques [23], frequently applied to visual tracking. In tracking, the goal is to estimate at time t the object state $\mathbf{x}_t$ (e.g. location and size) using observations $\mathbf{y}_{0:t}$ (object appearance in video frames). The computation requires a likelihood of a hypothesized state that gives rise to observations and a transition model that describes the way states are propagated between frames. Since the likelihood models in practical situations lead to intractable exact inference, approximation by Monte Carlo methods, also known as particle filtering, has been widely adopted. At each time step t, the prediction step involves sampling from the proposal distribution $p(\mathbf{x}_t|\mathbf{x}_{0:t-1}, \mathbf{y}_{0:t})$ of the current state $\mathbf{x}_t$ conditioned on the history of states $\mathbf{x}_{0:t-1}$ up to time $t-1$ and the history of observations $\mathbf{y}_{0:t}$ up to time t. The estimate is then computed during the update step based on the prediction and all observations. In detection, the sequence of probability distributions specifies a spatial order rather than a time order. The posterior distribution of the pose (state) of each anatomical structure is estimated based on all observations so far. The observations are features computed from image neighborhoods surrounding the anatomies. The likelihood of a hypothesized state that gives rise to observations is based on a deterministic model learned using a large annotated database of images. The transition model that describes the way the poses of anatomical structures are related is Gaussian. We will discuss several transition models specific to fetal head anatomies.

The computational speed and robustness of our system is increased by hierarchical processing. In detection, one major problem is how to effectively propagate detection candidates across the levels of the hierarchy. This typically involves defining a search range at a fine level where the candidates from the coarse level are refined. Incorrect selection of the search range leads to higher computational cost, lower accuracy, or drift of the coarse candidates towards incorrect refinements. The search range in our technique is part of the model that is

learned from the training data. The performance of our detection system is further improved by starting from structures that are easier to detect and constraining the detection of the other structures by exploiting spatial configurations. The difficulty of this strategy is selecting the order of detections such that the overall performance is maximized. Our detection schedule is designed to minimize the uncertainty of the detections. The Automatic Fetal Head and Brain measurement system automatically finds the standard visualization plane of a requested anatomy and displays the measurement value. The system is fast: eight structures and six measurements are displayed on a standard desktop computer within 14.7 or 6.9 seconds when using CPU or GPU, respectively. The average difference between ground truth and automatic measurements is below 2.0 mm and all measurements are within inter-user variability.

Compared to an earlier version of the work [24], this paper handles more structures (eight structures and six measurements vs. only three structures and three measurements) by introducing new anatomy-specific transition models. The need for a large number of observation and transition models is addressed by the generalization of the earlier architecture yielding the Integrated Detection Network (IDN). The experiments are performed on more data sets (2089 vs. only 884 in [24]) and more thorough evaluation uses two types of measurements and two baseline comparisons. Finally, the experiments are extended with a new section on clinical evaluations using an international panel of experts. Previous algorithm [10] handled only three structures and three measurements. They were computed at a single image resolution level which resulted in lower accuracy and higher computational cost.

The paper is organized as follows. We start by giving an overview of the anatomical measurements of fetal head and brain obtained in the second and third trimesters of pregnancy (Section II). We continue by reviewing background literature on detection and measurements in ultrasound images and relevant literature on multi-object detection in computer vision (Section III). The Automatic Fetal Head and Brain measurements algorithm is explained in Section IV. The evaluation in Section V focuses on qualitative and quantitative analysis of the AFHB system. Section VI summarizes clinical evaluations performed at several sites internationally. We will conclude the paper in Section VII.

II. Anatomical Measurements of Fetal Head and Brain Structures

Transabdominal sonography is the technique of choice to investigate the fetal Central Nervous System (CNS) during late first, second and third trimesters of gestation in low risk pregnancies [6]. Major components of this exam are visualizations and measurements of fetal head and brain structures in standardized planes shown in Figure 1.

During second and third trimesters, three standardized scan planes, *thalamic*, *ventricular*, and *cerebellar*, allow detection of most cerebral anomalies [25].

The *thalamic* plane is used for the biometry of the head, namely for measuring the biparietal diameter (BPD) and the occipitofrontal diameter (OFD). The biparietal diameter is measured outer-to-outer (i.e. across two outer boundaries of the skull). The head circumference (HC) is computed as an ellipse cirucmference using BPD and OFD measurements as ellipse axes. The major landmarks in the thalamic plane include the frontal horns of the lateral ventricles, the cavum septi pellucidi, the thalami, and the hippocampal gyruses [6]. The plane should not show any part of the cerebellum.

The *ventricular* plane is slightly craniad to the thalamic plane and it gives optimal visualization of the body and atrium of the lateral ventricles [25]. The ventricles are medially separated by the cavum septi pellucidi (CSP). To find the ventricular plane, a clinician needs to see cavum septi pellucidi, frontal horn, atrium, and choroids plexus. The plane is used to measure the size of lateral ventricles (LV). The best measurement is obtained by measuring the inner diameter (width) of the atrium (Figure 1). In the standard ventricular plane, only the hemisphere on the far side of the transducer is usually clearly visualized, as the hemisphere close to the transducer is frequently obscured by artifacts.

The *cerebellar* plane is obtained by slight posterior rotating the thalamic plane to show the cerebellar hemispheres. The cerebellum appears as a butterfly-shaped structure formed by the round cerebellar hemispheres [6]. Other landmarks include the cisterna magna, the thalamus, and the cavum septi pellucidi. The cisterna magna is a fluid-filled space posterior to the cerebellum [6]. The cerebellar plane is used to measure the antero-posterior diameter of the cisterna magna (CM) and the width of the cerebellar hemispheres (CER) [25].

The *median* (or *mid-sagittal*) plane shows the corpus callosum (CC) with all its components; the cavum septum pellucidi, the brain stem, pons, vermis, and posterior fossa [6]. The corpus callosum is a midline structure overlying the lateral ventricles that carries nerve fibers connecting the right and left hemispheres. The visualization of this plane is used to assess the anomalies and agenesis of the corpus callosum. There is no measurement associated with CC.

The lateral ventricle is filled with echogenic choroid plexus (CP). As with lateral ventricles, only the hemisphere on the far side of the transducer is usually clearly visualized. The plane is located craniad from the cerebellar plane. The visualization of this plane is useful for examining choroid plexus for cysts.

III. Prior work on Quantitative Ultrasound Analysis

In this section, we review previously published literature on the detection and segmentation of anatomical structures in ultrasound images. In addition to the challenges in echocardiography [22], we are facing new challenges in the obstetrics application: (1) The algorithm needs to handle several different structures in the brain, (2) the appearance of structures varies more dramatically across the data set, (3) relationships between structures should be explored to take advantage of the additional constraints, and (4) the algorithm must be very efficient. We start the review by a survey of works on detecting and measuring structures in 2D and 3D ultrasound images. We then examine a few papers on segmentation of ultrasound

In some countries, outer-to-inner measurement for BPD is more common.

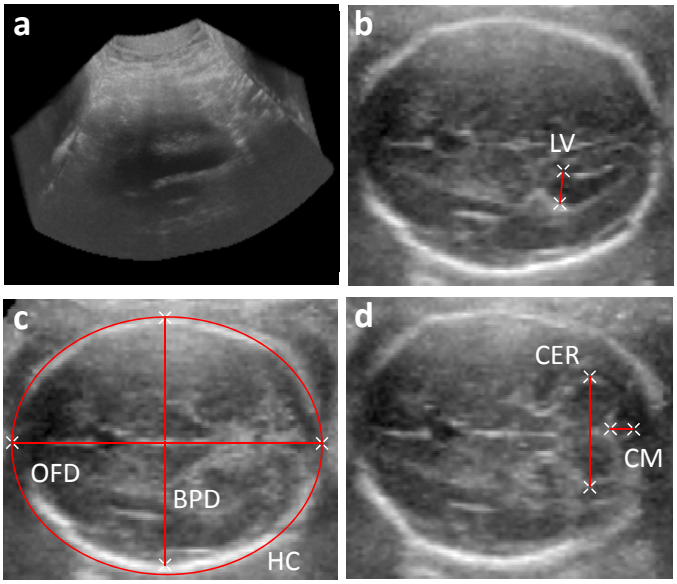

Fig. 2. Using the input volume (a), Automatic Fetal Head and Brain (AFHB) system provides the following measurements: (b) Lateral Ventricles (LV), (c) Cerebellum (CER) and Cisterna Magna (CM), and (d) Occipitofrontal Diameter (OFD), Biparietal Diameter (BPD), and Head Circumference (HC). In addition, the system provides automatic detection of median plane for visualization of Corpus Callosum (CC) and a plane for visualization of Choroid Plexus (CP), see Figure 1.

structures, which could be used to provide measurements. We also briefly review several algorithms that detect and segment brain structures in CT and MRI. Finally, we examine several related papers on multi-object detection from computer vision.

The literature on detecting and measuring fetal head and brain structures in 3D ultrasound images is limited. This is due to the above challenges posed by the ultrasound modality, by the variability of the fetal anatomical structures, and by relatively new use of 3D ultrasound in obstetrics practice. To capture this high variability, the most promising techniques use non-parametric appearance models obtained from large databases of annotated examples. Carneiro et al. [10] presented a system for detecting and measuring three structures: cerebellum, cisterna magna, and lateral ventricles. The algorithm uses a pose detector trained with Probabilistic Boosting Tree (PBT) [26] to find the structures and semi-local context to leverage prior knowledge of structure locations. However, only three structures (two standardized planes) are found and the algorithm operates on a single resolution level. This results in lower accuracy and higher computational cost. Yaqub et al. [7] trained random forest classifier on 10 volumes to classify voxels to belong to one of five classes: background, choroid plexus, posterior ventricle cavity, cavum septum pellucidum, and cerebellum. The results shows that discriminative techniques are useful for identifying structures in ultrasound volumes. However, the testing data set is small (only 10 volumes), the visualization plane must be found manually, and the technique does not produce automatic measurements. Pauly et al. [8] proposes to detect substantia nigra echogenicities in 3D transcranial ultrasound by a probabilistic model consisting of data and prior terms learned by random forests. The voxel classification uses mean intensity features and exploits the symmetry of the brain anatomy. The above papers [7], [8] identify and detect structures in ultrasound volumes of the fetus but do not focus on providing highly accurate measurements of the structures.

Several papers in the literature are concerned with detecting and measuring structures other than brain. The approaches focus on designing specialized features [11], [27] and use discriminative learning to build robust detectors [12], [28]. Rahmatullah et al. [28] automatically find stomach visualization plane in 3D ultrasound volumes by searching for stomach bubble and umbilical vein landmarks. The algorithm relies on AdaBoost training algorithm and detects the plane in 6 seconds on average with model trained on 2384 images. This computation time is lowered to under 1 second by using phase-based global features in the first level of the detector [11]. The algorithm only identifies an axis-aligned plane in a 3D scan and therefore does not find the plane orientation. Romeny et al. [27] find centers of ovarian follicles using the so-called winding number of the intensity singularity. The number is indicative of intensity minima at a given scale. It is computed as a path integral of the angular increment of the gradient direction vector in a closed neighborhood. Chen et al. [12] estimate the size and position of individual ovarian follicles by a probabilistic framework. In this framework, the parameter space is partitioned and clustered to efficiently search the hypotheses in a high dimensional space. The best position and size candidate is used to initialize robust 3D segmentation. Some of the above algorithms use discriminative learning to reach accurate results at acceptable computational speeds [12], [28], [29]. However, none of the techniques provides plane orientation for the visualization of the structures. In addition, it is not always straightforward how to extend these algorithms to the fetal head and brain structures.

Sometimes, the detection algorithms are used to initialize tracking [22] or segmentation [12] of anatomical structures in 3D ultrasound. The literature on (semi-) automatic segmentation is larger than literature on automatic detection and measurement (see survey by Noble et al. [30]). These techniques could, in principle, be used to provide measurements of the structures after they have been segmented. The most reliable algorithms are based on pixel-wise classifiers [7], machine learning [11], [12], and deformable models [13], [14]. Gooding et al. [9] propose a semi-automatic method for segmenting ovarian follicles. The method relies on the level-set framework to incorporate regional constraints and to deal with signal dropouts. The segmentation results are used to report volume measurements of the follicles. Hong et al. [13] propose a 3D left ventricle segmentation algorithm which uses a set of 2D slices and a discriminative classifier to find the ventricle boundary. The consistency of the slices across shapes is ensured by a non-rigid shape alignment step. Juang et al. [31] propose a graph-based technique for automatic segmentation of the left ventricle and atrium in 3D ultrasound volumes. The graph is constructed in a cylindrical coordinate space as determined by a central axis from the radial symmetry transform. In [32] textures extracted with Gabor filters are classified as belonging to prostate or background. The extracted

textures can also be used within a hierarchical deformable segmentation model [14] to produce 3D segmentation. All these segmentation techniques deal with the same challenges of the ultrasound images stated above, but our goal is different. Unlike segmenting a particular structure in its entirety, we are concerned with providing an accurate measurement. Therefore, we need to find the pose of the structure as robustly as possible and then automatically compute the measurement. Although requirements on measurement accuracy are very high, the process of detecting and measuring the structure is typically much faster and more robust than segmentation, where delineation of the whole boundary needs to be found.

There have been several methods in literature focused on detecting and segmenting fetal structures in 2D images. The early techniques relied on filtering, morphological operators, and hough transform [33]. These techniques tend to be slow and are typically designed for a specific anatomy which makes them difficult to generalize to new structures. Chalana et al. [34] describe a method for detecting the biparietal diameter and head circumference based on active contour model. Although the technique is real-time, it requires manual initialization. The major drawback of this approach is the lack of the appearance term to improve the robustness and accuracy. Zhang et al. [35] first automatically select the plane for measuring gestational sac from a 2D ultrasound video by a multi-scale AdaBoost classifier. The sac is then measured in real-time by an active contour model initialized from normalized cuts. Carneiro et al. [36] proposed a system for detecting and measuring several anatomies in 2D ultrasound images using the same underlying algorithm. The method learns to discriminate between the structures of interested and background via Probabilistic Boosting Tree classifier [26]. An efficient search technique makes the system run in under half second. The published 2D methods for detecting and measuring fetal anatomical structures often provide robust results at fast computational speeds. However, the detection and automatic measurements are harder in 3D. First, the 2D algorithms do not need to find the best measurement plane (this is done during the acquisition by the sonographer). Second, the search space of the structure of interest is much smaller and therefore the 2D algorithms can be fast. Finally, the 3D anatomy necessitates that the features and prior constraints are designed in 3D rather than 2D.

Magnetic Resonance Imaging (MRI) has been often used to image brain of adults but it is less frequent for the imaging of fetuses. Tu et al. [37] combined discriminative classifier based on probabilistic boosting tree (PBT) [26] for appearance modeling and a generative classifier based on principal component analysis (PCA) for shape modeling. The weights for these two terms learned automatically. The system takes 8 minutes to run to segment 8 brain structures. Anquez et al. [38] segment fetal eyes and skull bone content in MRI volumes. The eyes are first found by template matching. The eye locations are then used to position a mean shape model of the skull. The segmentation of the skull bone content is performed by graph cuts, first in a midsagittal plane and then in 3D. MRI fetal scanning is expensive and not approved for fetuses of gestational age below 20 weeks. Due to motion of the fetus during the scan, the motion correction must be applied [39]. Finally, more constrained scanning procedure and different imaging characteristics (higher signal-to-noise ratio, no signal drop outs, and no shadowing artifacts) make these techniques difficult to extend to the ultrasound domain.

Detecting multiple objects is studied from multiple aspects in computer vision literature. In our review, we focus on sampling, multi-resolution, and detection order selection problems. Many object detection algorithms [16], [26], [40] test a discrete set of object poses for an object presence with a binary classifier. Unlike these algorithms, that typically sample the parameter space uniformly, we sample from a proposal distribution [41] that focuses on regions of high probability. This saves computational time as fewer samples are required and increases robustness compared to the case, where the same number of samples would be drawn uniformly. Speedup can also come from using regression forests and a sparse set of samples in a volume [42]. However, these techniques have only been applied to detecting axis-aligned bounding boxes around organs in CT volumes and it is not straightforward how to obtain highly accurate measurements of the organs.

Multi-object detection techniques have focused on models that share features [43] or object parts [17]. This sharing results in stronger models, yet in recent literature, there has been a debate on how to model the object context in an effective way [44]. It has been shown that the local detectors can be improved by modeling the interdependence of objects using contextual [45]–[47] and semantic information [48]. The relationships between parts is often modeled as a Markov random field or conditional random filed [49]. This results in accurate detection of each individual part but at a high computational cost. Therefore, it is not possible to apply these techniques in an online system. Furthermore, the part-based model assumes the parts undergo articulate motion or non-rigid deformation which is not the case in fetal head and brain structures. In our Sequential Sampling framework, the interdependence between objects is modeled by a transition distribution, that specifies the "transition" of a pose of one object to a pose of another object. This way, we make use of the strong prior information present in medical images of human body. The important questions are how to determine the size of the context region (detection scale) and which objects to detect first in an optimal way.

Multi-scale algorithms usually specify a fixed set of scales with predetermined parameters of the detection regions [17], [50]. Choosing the scale automatically has the advantage since objects have different sizes and the size of the context neighborhood is also different. We propose a multi-scale scheduling algorithm that is formulated in the same way as the detection order scheduling.

The order of detection has been specified by maximizing the information gain computed before and after the detection measurement is taken [51], by minimizing the entropy of posterior belief distribution of observations [50], and by submodular maximization to guarantee optimal detection speed [52]. Our scheduling criterion is based on probability of states (object poses) within the ground truth region. Other measures could be used as well thanks to the flexible nature of the Sequential

Sampling framework.

IV. Sequential Sampling for Multi-Object Detection

The input for our system is an ultrasound volume containing the head of a fetus with a gestational age between 16 and 35 weeks. For fetuses older than 35 weeks, the ultrasound signal has a difficulty penetrating the skull. The brain structures are detected by a multi-object detection system as follows. The state (pose) of the modeled object s is denoted as $\boldsymbol{\theta}_s$ and the sequence of multiple objects as $\boldsymbol{\theta}_{0:s} = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_s\}$. In our case, $\boldsymbol{\theta}_s = \{\mathbf{p}, \mathbf{r}, \mathbf{s}\}$ denotes the position $\mathbf{p}$, orientation $\mathbf{r}$, and size $\mathbf{s}$ of the object s. The set of observations for object s are obtained from the image neighborhood V_s. The neighborhood V_s is specified by the coordinates of a bounding box within an 3-dimensional image V, $V : R^3 \rightarrow I$, where I is image intensity. The sequence of observations is denoted as $V_{0:s} = \{V_0, V_1, \ldots, V_s\}$. This is possible since there exists prior knowledge for determining the image neighborhoods $V_0, V_1, \ldots, V_s$. The image neighborhoods in the sequence $V_{0:s}$ might overlap and can have different sizes (Figure 3). An image neighborhood V_i might even be the entire volume V. The observations V_s with a likelihood $f(V_s|\boldsymbol{\theta}_s)$ describe the appearance of each object and are assumed conditionally independent given the state $\boldsymbol{\theta}_s$. The state dynamics, *i.e.* relationships between object poses, are modeled with an initial distribution $f(\boldsymbol{\theta}_0)$ and a transition distribution $f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{0:s-1})$. Note that here we do not use the first-order Markov transition $f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{s-1})$ often seen in visual tracking.

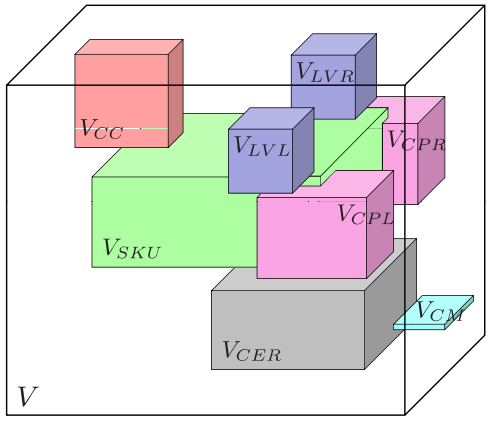

Fig. 3. In multi-object detection, the set of observations is a sequence of image patches $\{V_s\}$. The sequence specifies a spatial order of structures. The structures are detected in this order which is automatically determined.

The multi-object detection problem is solved by recursively applying *prediction* and *update* steps to obtain the posterior distribution $f(\boldsymbol{\theta}_{0:s}|V_{0:s})$. The prediction step computes the probability density of the state of the object s using the state of the previous object, $s-1$, and previous observations of all objects up to $s-1$:

$$f(\boldsymbol{\theta}_{0:s}|V_{0:s-1}) = f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{0:s-1}) f(\boldsymbol{\theta}_{0:s-1}|V_{0:s-1}). \quad (1)$$

When detecting object s, the observation V_s is used to compute the estimate during the update step as:

$$f(\boldsymbol{\theta}_{0:s}|V_{0:s}) = \frac{f(V_s|\boldsymbol{\theta}_s) f(\boldsymbol{\theta}_{0:s}|V_{0:s-1})}{f(V_s|V_{0:s-1})}, \quad (2)$$

where $f(V_s|V_{0:s-1})$ is the normalizing constant.

As simple as they seem these expressions do not have analytical solution in general. This problem is addressed by drawing m weighted samples $\{\boldsymbol{\theta}^j_{0:s}, w^j_s\}^m_{j=1}$ from the distribution $f(\boldsymbol{\theta}_{0:s}|V_{0:s})$, where $\{\boldsymbol{\theta}^j_{0:s}\}^m_{j=1}$ is a realization of state $\boldsymbol{\theta}_{0:s}$ with weight w^j_s.

In most practical situations, sampling directly from $f(\boldsymbol{\theta}_{0:s}|V_{0:s})$ is not feasible. The idea of importance sampling is to introduce a *proposal distribution* $p(\boldsymbol{\theta}_{0:s}|V_{0:s})$ which includes the support of $f(\boldsymbol{\theta}_{0:s}|V_{0:s})$.

In order for the samples to be proper [41], the weights are defined as

$$\tilde{w}^j_s = \frac{f(V_{0:s}|\boldsymbol{\theta}^j_{0:s}) f(\boldsymbol{\theta}^j_{0:s})}{p(\boldsymbol{\theta}^j_{0:s}|V_{0:s})}$$
$$w^j_s = \tilde{w}^j_s / \sum_{i=1}^{m} \tilde{w}^i_s. \quad (3)$$

Since the current states do not depend on observations from other objects then

$$p(\boldsymbol{\theta}_{0:s}|V_{0:s}) = p(\boldsymbol{\theta}_{0:s-1}|V_{0:s-1}) p(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{0:s-1}, V_{0:s}). \quad (4)$$

Note, that V_s was left out of the first term since the states in the sequence $\boldsymbol{\theta}_{0:s-1}$ do not depend on it. The states are computed as

$$f(\boldsymbol{\theta}_{0:s}) = f(\boldsymbol{\theta}_o) \prod_{j=1}^{s} f(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{0:j-1}). \quad (5)$$

Substituting (4) and (5) into (3), we have

$$\tilde{w}^j_s = \frac{f(V_{0:s}|\boldsymbol{\theta}^j_{0:s}) f(\boldsymbol{\theta}^j_{0:s})}{p(\boldsymbol{\theta}^j_{0:s-1}|V_{0:s-1}) p(\boldsymbol{\theta}^j_s|\boldsymbol{\theta}^j_{0:s-1}, V_{0:s})} \quad (6)$$
$$= \tilde{w}^j_{s-1} \frac{f(V_{0:s}|\boldsymbol{\theta}^j_{0:s}) f(\boldsymbol{\theta}^j_{0:s})}{f(V_{0:s-1}|\boldsymbol{\theta}^j_{0:s-1}) f(\boldsymbol{\theta}^j_{0:s-1}) p(\boldsymbol{\theta}^j_s|\boldsymbol{\theta}^j_{0:s-1}, V_{0:s})} \quad (7)$$
$$= \tilde{w}^j_{s-1} \frac{f(V_s|\boldsymbol{\theta}^j_s) f(\boldsymbol{\theta}^j_s|\boldsymbol{\theta}^j_{0:s-1})}{p(\boldsymbol{\theta}^j_s|\boldsymbol{\theta}^j_{0:s-1}, V_{0:s})}. \quad (8)$$

In this paper, we adopt the transition prior $f(\boldsymbol{\theta}^j_s|\boldsymbol{\theta}^j_{0:s-1})$ as the proposal distribution. Compared to the more general proposal, $p(\boldsymbol{\theta}^j_s|\boldsymbol{\theta}^j_{0:s-1}, V_{0:s})$, the most recent observation is missing. In practice, this does not pose a problem in our application since the predicted samples are near the likelihood peaks. The importance weights are then calculated as:

$$\tilde{w}^j_s = \tilde{w}^j_{s-1} f(V_s|\boldsymbol{\theta}^j_s). \quad (9)$$

In future, we plan to design more sophisticated proposal distributions to leverage relations between multiple objects during detection.

When detecting each object, the sequential sampling produces the approximation of the posterior distribution $f(\boldsymbol{\theta}_{0:s}|V_{0:s})$ using the samples from the detection of the previous object as follows:

1) Obtain m samples from the proposal distribution, $\boldsymbol{\theta}_s^j \sim p(\boldsymbol{\theta}_s^j|\boldsymbol{\theta}_{0:s-1}^j)$.
2) Reweight each sample according to the importance ratio

$$\tilde{w}_s^j = \tilde{w}_{s-1}^j f(V_s|\boldsymbol{\theta}_s^j). \tag{10}$$

Normalize the importance weights.
3) Resample the particles using their importance weights to obtain more particles in the peaks of the distribution. Finally, compute the approximation of $f(\boldsymbol{\theta}_{0:s}|V_{0:s})$:

$$f(\boldsymbol{\theta}_{0:s}|V_{0:s}) \approx \sum_{j=1}^{m} w_s^j \delta(\boldsymbol{\theta}_{0:s} - \boldsymbol{\theta}_{0:s}^j), \tag{11}$$

where δ is the Dirac delta function.

A. The Observation and Transition Models

The set of best instance parameters $\hat{\boldsymbol{\theta}}_s$ for each object s is estimated using the observations V_s

$$\hat{\boldsymbol{\theta}}_s = \arg\max_{\boldsymbol{\theta}_s} P(V_s|\boldsymbol{\theta}_s). \tag{12}$$

Let us now define a random variable $y_s \in \{-1, +1\}$, where $y_s = +1$ indicates the presence and $y_s = -1$ absence of the object s. To leverage the power of a large annotated dataset, we use a discriminative classifier (PBT [26]) to best decide between positive and negative examples of the object. PBT combines a binary decision tree with boosting, letting each tree node be an AdaBoost classifier. This way, the miss-classified positive or negative examples early on can still be correctly classified by children nodes. We can now evaluate the probability of an anatomy being detected as $P(y_s = +1|\boldsymbol{\theta}_s, V_s)$, which denotes posterior probability of the object presence with parameters $\boldsymbol{\theta}_s$ given observations V_s. This is a natural choice for the observation model in Eq. (12),

$$\hat{\boldsymbol{\theta}}_s = \arg\max_{\boldsymbol{\theta}_s} P(y_s = +1|\boldsymbol{\theta}_s, V_s) \tag{13}$$

In tracking, often a Markov process is assumed for the transition kernel $f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{0:s-1}) = f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{s-1})$, as time proceeds. However, this is too restrictive for multiple object detection. The best transition kernel might stem from an object different from the immediate precursor, depending on the anatomical context. In this paper, we use a pairwise dependency

$$f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{0:s-1}) = f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_j), \quad j \in \{0, 1, \ldots, s-1\}. \tag{14}$$

We model $f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_j)$ as a Gaussian distribution estimated from the training data. The statistical model captures spatial relationships between the structures while ignoring abnormal configurations that may be caused by brain malformations. During detection, the predictions are used as the best available estimates even for abnormal cases.

B. Integrated Detection Network (IDN)

Large number of observation and transition models creates a need for flexible architecture that would simplify algorithm design. To address this need, we propose Integrated Detection Network (IDN), which simplifies design, modification, tuning, and implementation of sophisticated detection systems [53]. As shown in Figure 4(left), IDN is a pairwise, feed-forward network. IDN consists of *nodes* that perform operations on the input *data* and produce zero or more output data. The operations, such as candidate sample detection, propagation, and aggregation, are only related to each other through data connections. This makes it possible to easily add new nodes and data types to an existing network. The design is a generalization of the Hierarchical Detection Network (HDN) [24], which only focused on detection nodes. The same network is used in both detection and training which enables rapid prototyping and algorithm evaluation. Furthermore, the basic network building blocks (such as rigid detector encapsulating position, orientation, and size detection) can be designed and interconnected into complex hierarchies. Such flexible design makes it easy to manage large-scale detection systems.

C. Detection Order Selection

Unlike a video, where the observations arise in a naturally sequential fashion, the spatial order in multi-object detection must be selected. The goal is to select the order such that the posterior probability $P(\boldsymbol{\theta}_{0:s}|V_{0:s})$ is maximized in the neighborhood region around the ground truth. Since determining this order has exponential complexity in the number of objects, we adopt a greedy approach. We first split the training data into two sets. Using the first set, we train all object detectors individually to obtain posterior distributions $f(\boldsymbol{\theta}_0|V_0), f(\boldsymbol{\theta}_1|V_1), \ldots, f(\boldsymbol{\theta}_s|V_s)$. The second set is used for order selection as follows.

We aim to build an Integrated Detection Network (IDN) from the order selection as illustrated in Figure 4(right). Suppose that we find the ordered detectors up to $s-1$, $\boldsymbol{\theta}_{(0)}, \boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(s-1)}$. We aim to add to the network the best pair $[s, (j)]$ (or feed-forward path) that maximizes the expected value of the following score $S[s, (j)]$ over both s and (j) computed from the second training set:

$$S[s, (j)] = \int_{\substack{\boldsymbol{\theta}_s \in \Omega(\tilde{\boldsymbol{\theta}}_s) \\ \boldsymbol{\theta}_{(0:s-1)} \in \Omega(\tilde{\boldsymbol{\theta}}_{(0:s-1)})}} f(\boldsymbol{\theta}_{(0:s-1)}|V_{(0:s-1)}) f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{(j)}) f(V_s|\boldsymbol{\theta}_s) d\boldsymbol{\theta}_s d\boldsymbol{\theta}_{(0:s-1)}, \tag{15}$$

where $\Omega(\tilde{\boldsymbol{\theta}})$ is the neighborhood region around the ground truth $\tilde{\boldsymbol{\theta}}$. The expected value is approximated as the sample mean of the cost computed for all examples of the second training data set. The order selection above is related to the learning structured predictors [19], [20] with the difference of selecting the predicted objects in a greedy fashion.

During hierarchical detection, larger object context is considered at coarser image resolutions resulting in robustness against noise, occlusions, and missing data. High detection accuracy is achieved by focusing the search in a smaller neighborhood at the finer resolutions. The resolution level and the size of the image neighborhoods $\{V_i\}$ can be selected using the same mechanism as the order selection by introducing additional parameters [24].

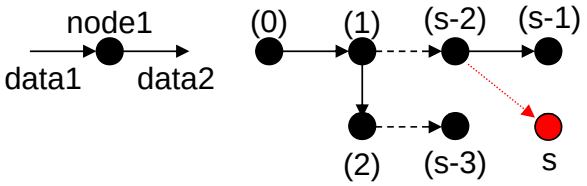

Fig. 4. Integrated Detection Network (IDN) consists of *nodes* that operate on *data* (left). Illustration of the Integrated Detection Network (IDN) and order selection (right). See text for details.

D. Anatomy-specific Transition Models

Sequential sampling in the detection of fetal head and brain structures is constrained by the anatomical domain present in the neurosonography scans. We take advantage of this constraint by designing two anatomy-specific transition models, one for the cisterna magna and one for lateral ventricles and choroid plexus.

The cisterna magna is found posterior to the cerebellum in the cerebellar plane. Both structures are measured in the same plane which reduces the search space of the automatic detection: the pose of the cisterna magna is predicted from the pose of the cerebellum *within* the cerebellar plane. We can write the transition model from (14) as

$$f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_j) = f(\boldsymbol{\theta}_{CM}|\hat{\boldsymbol{\theta}}_{CER}) = \mathcal{N}(\mu_{CER-CM}, \sigma^2_{CER-CM}), \tag{16}$$

where $\mu_{CER-CM} = \{\mathbf{p}_{CER-CM}, \mathbf{r}_{CER-CM}, \mathbf{s}_{CER-CM}\}$ and σ^2_{CER-CM} denote the mean and variance of the pose transition between cerebellum pose estimate $\hat{\boldsymbol{\theta}}_{CER}$ and cisterna magna. Furthermore, $\mathbf{p}_{CER-CM} = (p_x, p_y, 0)$ and $\mathbf{r}_{CER-CM} = (r_x, r_y, 0)$ due to the planar constraint. It is important to note, that the CER estimate $\hat{\boldsymbol{\theta}}_{CER}$ is computed first using all samples, rather than sequentially propagating them. This has two effects. First, CM samples are enforced to lie on the plane of the CER estimate. Second, the orientation of CM samples is within the CER estimate plane. As a result, this prediction model yields faster CM detection due to the reduced search subspace and more robust results due to the elimination of incorrect CER samples and enforced local search. Typically, 200 samples are used for position and orientation and 50 samples are used for size estimation.

It is possible to make use of the constrained scanning procedure to design transition models specific to LV and CP. The fetal head volume is typically scanned through the sphenoid phontanel and squamosal suture by finding the right angle along the side of the head. Following this procedure, the acquired scan shows the axial view and the variation of the scan orientation is constrained by the acquisition angle. The choroid plexus plane always have constrained orientation w.r.t. the scanning probe. Since only the hemisphere on the far side of the transducer can be clearly visualized (Section II), only LV measurement and CP visualization in this hemisphere is required. We take advantage of this constraint by training a separate detector for LV and CP in left and right hemispheres. The samples are propagated only to one side as determined by the orientation of the head from cerebellum candidates. The transition model assumes the following form:

$$f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_j) = f(\boldsymbol{\theta}_{LV}|\boldsymbol{\theta}_{CER}) = \begin{cases} \mathcal{N}(\mu_{CER-LVL}, \sigma^2_{CER-LVL}) \\ \arccos(\mathbf{q}^j_{CER} \cdot \mathbf{q}_r) > 0 \\ \quad \text{for at least half samples} \\ \mathcal{N}(\mu_{CER-LVR}, \sigma^2_{CER-LVR}) \\ \text{otherwise,} \end{cases} \tag{17}$$

where $\mu_{CER-LVL}$, $\mu_{CER-LVR}$ and $\sigma_{CER-LVL}$, $\sigma_{CER-LVR}$ denote the mean and variance of the pose transition between cerebellum and left and right lateral ventricle. The orientation $\mathbf{q}^j_{CER} = (q^j_x, q^j_y, q^j_z)$ specifies the orientation of a cerebellum sample j. The reference orientation $\mathbf{q}_r = (0, 1, 0)$ indicates the volume orientation vector. Due to the constrained scanning procedure, the angle between $\mathbf{q}_{CER}$ and $\mathbf{q}_r$ will be close to 0 or 180 degrees depending on whether the brain was scanned through the left or right side of the head. In practice, only one of the transition models is used in each volume. This way, the candidates are always sampled in the hemisphere further away from the probe which increases the robustness and speed of the lateral ventricle detection. Similar procedure is followed for CP transition model.

	Antares	S2000	Total
CER, CM, LV	884	1205	2089
HC, BPD, OFD	365	1206	1571
CC, CP	0	1193	1193
Total (all structures)	3747	9619	13366

TABLE I
ANNOTATION COUNTS FOR EACH STRUCTURE IN IMAGES FROM SIEMENS ANTARES AND SIEMENS S2000 SYSTEMS.

V. EXPERIMENTAL EVALUATION

The AFHB algorithm is evaluated in terms of the ability to detect the correct anatomical structures (Section V-A) and in terms of providing accurate measurements of these structures (Section V-B). The next section (Section VI) summarizes clinical evaluations using feedback from experts in ultrasound obstetrics exam.

We use a total of 2089 fetal head volumes with sizes ranging from $94\times75\times125$ mm to $279\times173\times221$ mm and the size average of $186\times123\times155$ mm. The volumes were acquired using Siemens Antares and Siemens S2000 ultrasound systems. The data sets were converted from acoustic to Cartesian coordinate system and resampled to 1 mm^3 resolution. The gestational ages ranged from 16 weeks to 35 weeks. The annotation counts for each structure are summarized in Table I. Overall, 13366 structures were annotated by an experienced sonographer with higher counts for S2000 volumes. Out of the total of 2089 volumes, 1982 were used for training and 107 for testing. To further increase the number of annotations, we took advantage of the symmetry in the anatomy of the head and a standardized position of the head in the scan (axial acquisition through the squamosal suture). The data (and annotations) were flipped along X, Z, and XZ axes, which resulted in a total of 8356 volumes and 53464 annotations.

All structures were annotated by finding the accurate plane and drawing the measurement line according to the guidelines of the International Society of Ultrasound in Obstetrics & Gynecology [6] (see Section II for an overview). Since the corpus callosum does not have a measurement associated with it, for the purposes of training a detector, we annotated this structure as follows. The annotation line was drawn in the *median* plane from the bottom of the genu inside the body of corpus callosum. Similarly, the choroid plexus is annotated by extending the annotation line along its longest axis in the plane where the choroid plexus is most visible (see Figure 10). The annotation planes together with the measurement lines define uniquely the poses. The line center, the line orientation with the annotation plane, and the line length define the position $\mathbf{p}$, orientation $\mathbf{r}$, and size $\mathbf{s}$ parameters of each pose. Therefore, the position of the annotation plane is defined by the line center and the orientation of the plane is defined by the line orientation. The 3D size parameter of the pose is extended by a fixed ratio to increase the context (the scaling ranges from 1.2 to 10.0 depending on the resolution and structure size). Since the BPD and OFD measurements are in the same plane, they are combined into a single pose. The 3D size parameter of the pose then consists of (OFD length $\times$ BDP length $\times$ *depth*). The *depth* was conveniently chosen as the BPD length. The poses are used to train a detector for each structure. The detected poses are mapped back to measurement lines and visualization planes for display.

A. IDN for Fetal Head and Brain Measurements

The structure poses defined in the previous section are used to train a Integrated Detection Network (IDN) (Section IV-B) on the training data set. The resulting network showing all structures at three resolution levels is shown in Figure 5. All-in-all, the network contains 10 detectors, 54 classifiers, and a total of 45 IDN nodes. The IDN nodes consists of position (10), orientation (9), and size detectors (8), as well as candidate prediction (7), splitting / joining (4), aggregation (6), and data handling (1). There is one detector for each of the eight structures with combined skull detector (SKU) for OFD, BPD, and HC, and there are two additional resolution levels for CER. LV and CP have a separate detector for left and right side (Section IV-D). Each detector consists of a classifier trained to detect position, position+orientation, and position+orientation+size (10 detectors $\times$ 3 = 30 classifiers). Each classifier is also bootstrapped to increase robustness against false positives [24], [54] (30 $\times$ 2 = 60). This is done by training the second classifier using the samples processed by the first classifier. There is no classifier for CER 4 mm orientation, size, and CER 2 mm size since the low resolution does not provide enough detail to detect these reliably [53] (60 - 6 = 54 classifiers). Additional nodes comprise of candidate aggregation and prediction [53].

Our first experiment tests the accuracy of correctly detecting the fetal head and brain structures. This is done by running the detector on each of the 107 unseen volumes and comparing the pose (position, orientation, and size) of each detected structure with the pose of the corresponding annotated structure. The orientation error was calculated by angular distance (using quaternion representation). For CC and CP, the orientation error only considers normal of the measurement plane. The size error was computed as the Euclidean distance between two 3D points, where the point vectors represent sizes of the detected and annotated structures.

	position [mm]	ort. [deg]	size [mm]
CER	1.75 $\pm$ 0.92	5.26 $\pm$ 2.33	3.56 $\pm$ 2.85
CM	1.95 $\pm$ 0.93	5.75 $\pm$ 2.20	3.02 $\pm$ 2.00
LV	1.72 $\pm$ 0.93	11.06 $\pm$ 6.12	3.51 $\pm$ 2.41
SKU	2.00 $\pm$ 0.89	7.03 $\pm$ 2.95	4.27 $\pm$ 2.85
CC	1.89 $\pm$ 0.96	9.46 $\pm$ 5.47	4.23 $\pm$ 3.73
CP	1.83 $\pm$ 0.78	11.29 $\pm$ 5.44	5.59 $\pm$ 3.86

TABLE II
AVERAGE POSE DETECTION ERRORS FOR EACH TRAINED DETECTOR APPLIED ON UNSEEN VOLUMES. STATISTICS WERE COMPUTED TO SHOW ERRORS IN POSITION, ORIENTATION, AND SIZE OF THE BOUNDING BOX. THE ERRORS TEND TO BE HIGHER WHEN THERE IS AN UNCERTAINTY IN ANNOTATION (E.G. SKU POSITION AND LV ORIENTATION). THE ERRORS TEND TO BE LOWER FOR STRUCTURES THAT ARE CLEAR TO ANNOTATE (E.G. CER ALL MEASURES AND CM ORIENTATION).

The quantitative analysis of the pose detection errors is in Table II. Mean of the 95% smallest errors was computed by comparing the detected locations to manual labeling. The position error is lower than 2 mm on all structures but skull (SKU) with standard deviation below 1 mm. The skull is the largest structure and it is hard to determine the skull center accurately even during annotation. The orientation error is lowest for CER and CM since there is no ambiguity in annotating these structures. The error is larger for LV but still acceptable as seen from the clinical evaluations (Section VI). Annotating LV is challenging due to small size and difficulty in finding the inner diameter of the atrium accurately [6]. Annotation of CP is also challenging since it is a banana-shaped structure and finding the best visualization plane can be hard. The size errors tend to be larger for larger structures which is in agreement with the level of accuracy that can be achieved when measuring these structures. Overall, the detection errors are lower when compared to [24] and [10]. These papers report position detection error above 2 mm for CER, CM, and LV.

B. Automatic Measurement

Our second analysis is focused on providing automatic measurement of fetal head and brain structures. The measurement value is shown along with the visualization of the measurement plane to the obstetrician during ultrasound exam. The measurement line is obtained directly from the detected pose. The position and orientation parameters define the center and orientation of the measurement line, respectively. The size parameter defines the measurement line length.

We computed the error in two different ways. The first value, *length error*, was computed as a difference between lengths of the automatic measurement line and the annotation line. This measure therefore ignores the inaccuracies in the structure position and plane orientation. Evaluating this

Large failures have low probability values and are deferred to manual measurement.

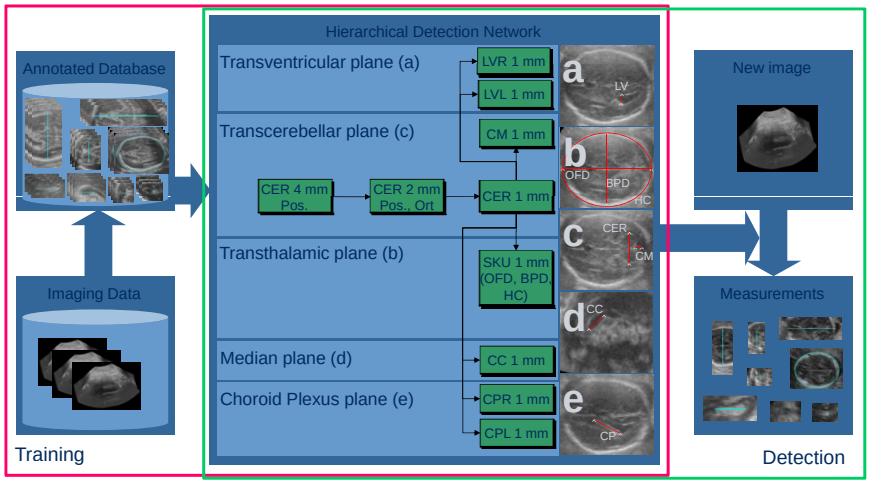

Fig. 5. During training, database of annotated images is used to train the Integrated Detection Network (IDN) at resolutions 4, 2, and 1 mm. During detection, the IDN is used to automatically provide measurements on a new image. The arrows in the IDN diagram indicate the interdependency and detection order of brain structures: Cerebellum (CER), Cisterna Magna (CM), Lateral Ventricles (LVL - left, LVR - right), Corpus Callosum (CC), and Choroid Plexus (CPL - left, CPR - right). The detection of Occipitofrontal Diameter (OFD), Biparietal Diameter (BPD), and Head Circumference (HC) is performed by the skull (SKU) detector.

error is useful since it determines the impact of using the automatically determined measurements in clinical practice. The second value, *plane error*, was computed as a maximum distance between corresponding line end-points (see Figure 6). This measure incorporates differences between detected and annotated positions and plane orientations.

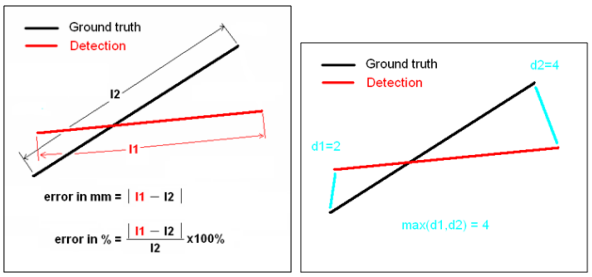

Fig. 6. Error measures computed to evaluate automatic measurements. The *length error* is computed as a difference between lengths of the automatic measurement and manual annotation. The *plane error* is computed as a maximum distance between corresponding line end-points. The latter also considers inaccuracies in structure position and plane orientation.

The evaluation of the automatic measurement errors is in Table III and Figure 7. The values for OFD, BPD, and HC were all computed from the detected skull (SKU) pose. The Head Circumference (HC) is computed as an ellipse circumference given major and minor axes as provided by OFD and BPD. The HC *plane error* is computed as a maximum of OFD and BPD plane error for each measurement. The table does not show errors for CC and CP since there is no clinical interest in the measurement of these structures. The *length error* is shown in the second column of the table. The error tends to be larger for larger measurements (e.g. HC and OFD). The average measurement lengths for all structures is in the seventh column. The third column shows the error as a percentage w.r.t. ground truth length. Small structures have higher percentage errors since even small inaccuracies have larger effect compared to the measurements of these structures. The *plane error* is shown in the fourth column. The values are higher than for the *length error* since this error measure takes into account inaccuracies in the plane orientation. The IDN plane errors are consistently lower when compared to the results when independently detecting structures at 1 mm resolution (shown in fifth column) and when detecting structures with a hierarchy of 4, 2, and 1 mm resolution detectors (sixth column). The plane errors for the hierarchical detection are lower than for the single-resolution detection. Comprehensive evaluation in Figure 7 shows the accuracy on all of the testing cases.

Our next evaluation includes visual comparison of different results along with the length and plane errors. Figure 8 shows example results for CER, CM, and LV. Figure 9 shows example results for OFD, BPD, and HC. Finally, Figure 10 shows results for CC and CP. Length and plane errors are reported for each case (note, that larger plane error does not necessarily mean that the length error will also be large as can be seen from the way they are computed (Figure 6). For each structure, two cases are shown: one case with *lower than average* plane error and one case with *higher than average* plane error as compared to the overall statistics (Table III). The purpose of this examination is three-fold. First, different scans are selected for each measurement which shows the high variability in the dataset. Second, the overview gives a clear idea how measurement errors correlate with the visual appearance of the respective planes and all structures that should be present according to the scanning guidelines [6]. Third, the larger plane errors do not necessarily mean that the scan would be rejected for diagnostic purposes – they are still acceptable for assessing fetal health, size, and growth. This is because the automatic measurements are within inter-user variability as we will show in the next section.

The average running times of the entire system (including data loading) is 45.2 seconds on computer with Intel Core 2 Duo, 2.66 GHz processor. When using GPU (nVidia

	length err.[mm]	length err.[%]	plane err.[mm]	plane err. 1 [mm]	plane err. 421 [mm]	GT length [mm]
CER	1.37 ± 1.10	3.96 ± 2.88	2.81 ± 1.24	2.84 ± 1.19	-	32.65 ± 6.60
CM	0.87 ± 0.58	15.29 ± 10.41	2.30 ± 0.96	4.87 ± 7.35	2.38 ± 1.21	5.70 ± 1.35
LV	1.01 ± 0.70	17.75 ± 13.19	2.21 ± 0.98	2.65 ± 1.51	2.28 ± 1.22	5.88 ± 1.31
OFD	2.31 ± 1.76	2.61 ± 1.93	5.69 ± 1.86	5.85 ± 2.33	5.91 ± 2.30	86.47 ± 11.81
BPD	0.94 ± 0.68	1.30 ± 0.88	4.74 ± 1.97	5.06 ± 2.33	5.03 ± 2.19	70.47 ± 11.24
HC	4.06 ± 2.82	1.61 ± 1.09	6.19 ± 1.95	6.65 ± 2.34	6.60 ± 2.33	247.57 ± 36.13

TABLE III

AVERAGE MEASUREMENT ERRORS FOR EACH TRAINED DETECTOR APPLIED ON UNSEEN VOLUMES. THE 2ND COLUMN SHOWS THE *length error* AND THE 4TH COLUMN SHOWS THE *plane error* COMPUTED ACCORDING TO FIGURE 6. THE PLANE ERROR IS LARGER OVERALL SINCE IT CONSIDERS INACCURACIES IN THE PLANE ORIENTATION. THE 3RD COLUMN SHOWS THE LENGTH ERROR AS A PERCENTAGE W.R.T. GROUND TRUTH LENGTH. THE 5TH AND 6TH COLUMN SHOW RESULTS OF DETECTING EACH STRUCTURE INDEPENDENTLY AT 1 MM RESOLUTION AND WITH A HIERARCHY OF 4, 2, AND 1 MM RESOLUTIONS, RESPECTIVELY. THE 7TH COLUMN SHOWS THE AVERAGE MEASUREMENT LENGTHS COMPUTED FROM GROUND TRUTH (GT). SEE THE TEXT FOR A DISCUSSION OF THESE RESULTS.

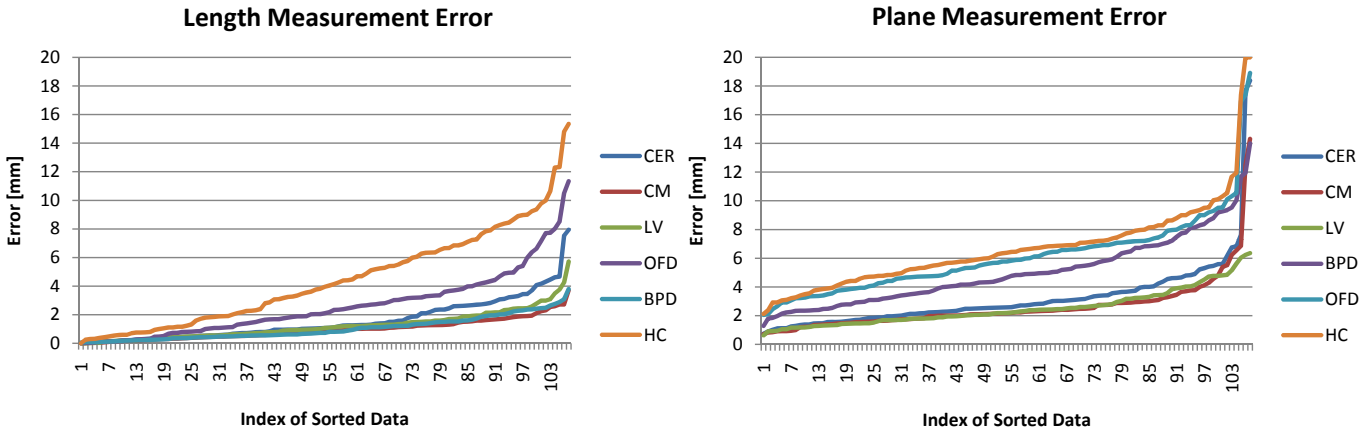

Fig. 7. Sorted length measurement errors (left) and plane measurement errors (right) computed for all structures using measures from Figure 6. The gradual increase of errors shows that the measurements are accurate for all structures with only a few outliers (towards the right of the figures). The largest error is for Head Circumference (HC) since it is the longest measurement (Table III. The plane measurement errors are larger since they also account for errors in the detection of the measurement plane.

GeForce 9800 GT) on the same machine, the running time is 6.9 seconds. The GPU implementation includes detectors (feature extraction and classification) for each structure. On a computer with Intel Six-core processor w. hyper threading, 2.40 GHz, the running time is 14.7 seconds using IDN network (Figure 5), 23.6 seconds using a hierarchy of 4, 2, and 1 mm resolutions for each structure independently, and 35.8 seconds using 1 mm resolution for each structure.

VI. CLINICAL EVALUATION

The AFHB algorithm was evaluated in a clinical setting. Total of 10 experts (doctors and sonographers) in ultrasound obstetrics exam were recruited from Europe, Asia, and United States. The results of the fully automatic algorithm are compared w.r.t. the experts and quantitatively analyzed in the next sections.

The following evaluation protocol was followed for each expert. The AFHB prototype system was installed on a laptop computer. All experts used the same set of 10 volumes of different quality and varying degree of imaging artifacts. The evaluation started by loading the first volume and running the AFHB algorithm. The user then selected the first structure which triggered automatic navigation to the measurement plane and automatic placement of the calipers to indicate the measurement. The user then had a choice to accept the automatic measurement and proceed to the next structure or manually correct the plane and the measurement result. This was done by using a mouse to navigate to the new measurement plane and manually placing the calipers to indicate the measurement. This process was repeated for all structures. Finally, the corrected results were saved such that the next volume can be processed. The evaluation focused on the measurements and therefore Corpus Callosum and Choroid Plexus planes were excluded.

Statistical analysis of the results is performed using modified Williams index [34]. Williams index measures an agreement of a reference user (e.g. automated system) with the joint agreement of other users. This is done by evaluating how much the reference user agrees with the other users compared to the agreement of the other users among themselves. Let us represent the database of measurements as $x_{i,j}$, $i \in 0,\dots n$, $j \in 1,\dots,N$, where i is a user index and j is case (image) index. The automated measurement is denoted by $i = 0$ and the users by $i \in \{1,\dots,n\}$. The index is computed for each structure s as

$$I_s = \frac{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{D_{0,j}}}{\frac{2}{n(n-1)}\sum_j\sum_{j':j'>j}\frac{1}{D_{j,j'}}}, \qquad (18)$$

where $D_{j,j'} = \frac{1}{N}\sum_{i=j}^{N} e(x_{i,j}, x_{i,j'})$ with e being the measurement error. The index is computed for both point and

length measurement error.

A confidence interval for the Williams index is computed using a jackknife non-parametric sampling technique [55]:

$$I_{s(\cdot)} \pm z_{0.95} se, \tag{19}$$

where $z_{0.95} = 1.96$ is the 95th percentile of the standard normal distribution. The jackknife estimate of the standard error is given by

$$se = \left\{ \frac{1}{N-1} \sum_{i=1}^{N} [I_{s(i)} - I_{s(\cdot)}]^2 \right\}^{1/2}, \tag{20}$$

where $I_{s(\cdot)} = \frac{1}{N} \sum_{i=j}^{N} I_{s(i)}$. The sampling procedure works by leaving out image i out of the computation of $D_{j,j'}$ and computing the Williams index $I_{s(i)}$ as in Eq. (18) for the remaining $N-1$ images. Reliable reference measurements have the value of the index close to 1.

A. Inter-user Variability

In this section, we analyze inter-user variability using evaluation protocol described in the previous subsection. Average length and plane errors were computed for each of the 10 users and each brain structure. The computation uses average user as ground truth and follows description from V-B and Figure 6. The results for length errors are summarized in Table IV and Figure 11(left) and results for plane errors are in Table V and Figure 11(right). Comparing to the results of automatic measurement using length and plane errors in Table III, it can be seen that the automatic results are within range of the inter-user variability. This demonstrates the high accuracy of the fully automatic measurement system. Tables IV and V also show the result of AFHB automatic measurement compared to the average user. These errors are lower than those in Table III since the experts were presented AFHB result for acceptance or correction (see evaluation protocol description in Section VI. Some users have larger errors than others. For example, users 4 and 6 have larger plane errors (Table V) than other users. This typically means that the users did not strictly follow the guidelines of the International Society of Ultrasound in Obstetrics & Gynecology [6] when adjusting the measurements. The plots in Figure 11 show the ranges of plane and length measurement errors (minimum and maximum) computed w.r.t. average user. The plots also show the deviation of each user w.r.t. initial AFHB result. It is clear that from these plots that the user variability can be quite large, especially for the longer measurements, and that the average changes to the results provided by AFHB system are small. All experts stated that the changes to the AFHB results are not clinically significant.

The Williams index for each measurement is shown in Table VI. The values of the confidence interval are close to 1 for all measurements. This represents a high level of agreement of the automatic measurements w.r.t. joint agreement of the users when compared to the agreement of the users among themselves.

User	CER	CM	LV	BPD	OFD	HC
1	1.66	0.73	0.62	1.75	1.73	3.39
2	1.70	1.03	0.80	1.65	2.56	6.28
3	1.38	0.77	0.67	2.58	1.58	3.14
4	1.14	0.83	0.59	1.35	2.33	4.80
5	1.43	0.84	0.57	1.35	1.83	4.19
6	1.56	1.81	0.77	1.23	2.88	5.19
7	1.35	1.08	0.73	1.53	1.94	3.46
8	1.51	1.14	0.77	1.54	2.41	4.34
9	1.33	1.08	0.63	1.41	1.63	3.36
10	1.50	0.85	0.58	1.46	1.99	3.76
Avg.	1.46	1.02	0.67	1.59	2.09	4.19
AFHB	1.35	0.62	0.53	1.04	1.85	3.99

TABLE IV
LENGTH ERROR (IN MM) COMPUTED ACCORDING TO FIGURE 6 FOR EACH OF THE 10 USERS AND STRUCTURE MEASUREMENT. AVERAGE ERROR WAS COMPUTED FOR EACH USER W.R.T. ALL OTHER USERS. THE LAST TWO ROWS SHOW THE AVERAGE INTER-USER VARIABILITY AND AVERAGE DEVIATION FROM THE INITIAL AFHB AUTOMATIC MEASUREMENT. THE LENGTH ERRORS COMPUTED FROM AUTOMATIC MEASUREMENTS (TABLE III) ARE WITHIN THE INTER-USER VARIABILITY.

User	CER	CM	LV	BPD	OFD	HC
1	2.57	1.62	1.84	6.93	6.45	6.58
2	3.27	2.49	1.94	5.94	7.42	7.51
3	2.93	1.89	1.69	5.03	5.09	5.23
4	2.97	1.90	2.64	8.81	16.35	16.36
5	2.56	1.64	1.63	4.45	5.14	5.31
6	2.59	2.93	1.95	6.36	9.82	10.34
7	2.92	2.41	2.23	4.55	5.34	5.56
8	3.24	2.54	2.18	5.01	5.67	5.74
9	2.42	2.00	1.86	4.51	5.39	5.76
10	2.59	1.79	1.76	4.62	5.35	5.43
Avg.	2.81	2.12	1.97	5.62	7.20	7.38
AFHB	1.74	1.87	1.32	3.23	5.35	5.39

TABLE V
PLANE ERROR (IN MM) COMPUTED ACCORDING TO FIGURE 6 FOR EACH OF THE 10 USERS AND STRUCTURE MEASUREMENT. AVERAGE ERROR WAS COMPUTED FOR EACH USER W.R.T. ALL OTHER USERS. THE LAST TWO ROWS SHOW THE AVERAGE INTER-USER VARIABILITY AND AVERAGE DEVIATION FROM THE INITIAL AFHB AUTOMATIC MEASUREMENT. THE LENGTH ERRORS COMPUTED FROM AUTOMATIC MEASUREMENTS (TABLE III) ARE WITHIN THE INTER-USER VARIABILITY.

VII. CONCLUSION

Wider acceptance of 3D ultrasound systems will be possible if they are easy to use, operator independent, and make the examination faster [4]. The Automatic Fetal Head and Brain (AFHB) measurement system proposed in this paper helps address this need. This unique system provides dramatic workflow improvements by reducing time required to measure anatomical structures, by reducing user variance for all measurements, and by increasing measurement accuracy. This is achieved by employing a sequential sampling model which makes it possible to use fewer samples of the structure pose and formally extend the class of binary classification algorithms [16], [26] to multiple structures. This saves computational time and increases accuracy since the samples are taken from the regions of high probability of the posterior distribution. This process is applied across a multi-resolution hierarchy when detecting one structure, but also when predicting pose parameters based on other structures. Such system is capable of resolving inherent challenges in obstetrics ultrasound imagery, such as speckle noise, signal

	WI length	95% CI length	WI plane	95% CI plane
CER	1.08	(0.94, 1.27)	1.61	(1.49, 1.77)
CM	1.65	(1.37, 2.05)	1.13	(0.94, 1.40)
LV	1.27	(1.10, 1.48)	1.49	(1.30, 1.75)
BPD	1.53	(1.28, 1.88)	1.74	(1.61, 1.88)
OFD	1.13	(0.96, 1.35)	1.35	(1.29, 1.41)
HC	1.37	(1.31, 1.43)	1.05	(0.88, 1.29)

TABLE VI
WILLIAMS INDEX (WI) AND CONFIDENCE INTERVALS (CI) COMPUTED FOR LENGTH AND PLANE MEASUREMENT ERRORS. THE VALUES ARE CLOSE TO 1, WHICH INDICATES AGREEMENT OF THE AUTOMATIC MEASUREMENT WITH THE USERS.

drop-out, strong shadows produced by the skull, and large intra-class variation caused by differences in gestational age. The system runs in 6.9 seconds on a GPU compatible with state-of-the-art ultrasound systems which makes it suitable for clinical use. All structures are accurately detected within inter-user variability.

The described framework opens up several possible avenues of future research. One area we are particularly interested in is how to include dependence on multiple objects at each detection stage. This will result in a stronger geometrical constraint and therefore improve performance on objects that are difficult to detect by exploiting only the pairwise dependence. This will also help when extending the AFHB system to other anatomical structures and measurements in routine ultrasound examinations.

REFERENCES

[1] A. B. Caughey, A. Ahsan, and L. M. Hopkins, *Obstetrics & gynecology*. Lippincott Williams & Wilkins, 2006.

[2] B. R. Benacerraf, "Three-dimensional fetal sonography; use and misuse," *J Ultrasound Med*, vol. 21, no. 10, pp. 1063–1067, 2002.

[3] L. Chan, T. Fung, T. Leung, D. Sahota, and T. Lau, "Volumetric (3D) imaging reduces inter- and intraobserver variation of fetal biometry measurements," *Ultrasound in Obstetrics and Gynecology*, vol. 33, no. 4, pp. 447–452, 2009.

[4] L. F. Goncalves, W. Lee, J. Espinoza, and R. Romero, "Three- and 4-dimensional ultrasound in obstetric practice: Does it help?" *J Ultrasound Med*, vol. 24, no. 3, pp. 1599–1624, 2005.

[5] B. Benacerraf, T. Shipp, and B. Bromley, "How sonographic tomography will change the face of obstetric sonography: A pilot study," *J Ultrasound Med*, vol. 24, no. 3, pp. 371–378, 2005.

[6] The International Society of Ultrasound in Obstetrics and Gynecology, "Sonographic examination of the fetal central nervous system: Guidelines for performing the basic examination and the fetal neurosonogram," *Ultrasound in Obstetrics and Gynecology*, vol. 29, no. 1, pp. 109–116, 2007.

[7] M. Yaqub, R. Napolitano, C. Ioannou, A. Papageorghiou, and J. Noble, "Automatic detection of local fetal brain structures in ultrasound images," in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, may 2012, pp. 1555 –1558.

[8] O. Pauly, S.-A. Ahmadi, A. Plate, K. Boetzel, and N. Navab, "Detection of substantia nigra echogenicities in 3d transcranial ultrasound for early diagnosis of parkinson disease," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2012*, ser. Lecture Notes in Computer Science, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds. Springer Berlin Heidelberg, 2012, vol. 7512, pp. 443–450.

[9] M. J. Gooding, S. Kennedy, and J. A. Noble, "Volume segmentation and reconstruction from freehand three-dimensional ultrasound data with application to ovarian follicle measurement," *Ultrasound in Medicine & Biology*, vol. 34, no. 2, pp. 183 – 195, 2008.

[10] G. Carneiro, F. Amat, B. Georgescu, S. Good, and D. Comaniciu, "Semantic-based indexing of fetal anatomies from 3-D ultrasound data using global/semi-local context and sequential sampling," in *Proc. CVPR*, Anchorage, AK, 24–26 Jun. 2008.

[11] B. Rahmatullah, A. Papageorghiou, and J. Noble, "Integration of local and global features for anatomical object detection in ultrasound," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2012*, ser. Lecture Notes in Computer Science, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds. Springer Berlin / Heidelberg, 2012, vol. 7512, pp. 402–409.

[12] T. Chen, W. Zhang, S. Good, K. Zhou, and D. Comaniciu, "Automatic ovarian follicle quantification from 3d ultrasound data using global/local context with database guided segmentation," in *Computer Vision, 2009 IEEE 12th International Conference on*, 29 2009-oct. 2 2009, pp. 795 –802.

[13] W. Hong, B. Georgescu, X. Zhou, S. Krishnan, Y. Ma, and D. Comaniciu, "Database-guided simultaneous multi-slice 3d segmentation for volumetric data," in *Computer Vision - ECCV 2006*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer Berlin / Heidelberg, 2006, vol. 3954, pp. 397–409.

[14] D. Shen, Y. Zhan, and C. Davatzikos, "Segmentation of prostate boundaries from ultrasound images using statistical shape model," *Medical Imaging, IEEE Transactions on*, vol. 22, no. 4, pp. 539 –551, april 2003.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, vol. 1, 2005, pp. 886–893.

[16] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comp. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[17] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[18] D. Crandall, P. Felzenszwalb, and D. Huttenlocher, "Spatial priors for part-based recognition using statistical models," in *Proc. CVPR*, vol. 1, 2005, pp. 10–17.

[19] T. Joachims, T. Hofmann, Y. Yue, and C.-N. J. Yu, "Predicting structured objects with support vector machines," *Commun. ACM*, vol. 52, no. 11, pp. 97–104, 2009.

[20] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin, "Learning structured prediction models: a large margin approach," in *Proceedings of the 22nd International Conference on Machine Learning (ICML)*. New York, NY, USA: ACM, 2005, pp. 896–903.

[21] B. Sapp, A. Toshev, and B. Taskar, "Cascaded models for articulated pose estimation," in *Proc. Eleventh ECCV*, 2010, pp. 406–420.

[22] L. Yang, B. Georgescu, Y. Zheng, Y. Wang, P. Meer, and D. Comaniciu, "Prediction based collaborative trackers (PCT): A robust and accurate approach toward 3D medical object tracking," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 11, pp. 1921 –1932, nov. 2011.

[23] A. Doucet, N. D. Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*. Birkhäuser, 2001.

[24] M. Sofka, J. Zhang, S. Zhou, and D. Comaniciu, "Multiple object detection by sequential Monte Carlo and hierarchical detection network," in *Proc. CVPR*, San Francisco, CA, 13–18 Jun. 2010.

[25] E. E. Sauerbrei, K. T. Nguyen, and R. L. Nolan, *A Practical Guide to Ultrasound in Obstetrics and Gynecology*. Lippincott-Raven Publishers, 1998.

[26] Z. Tu, "Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering," in *Proc. ICCV*, vol. 2, 2005, pp. 1589–1596.

[27] B. ter Haar Romeny, B. Titulaer, S. Kalitzin, G. Scheffer, F. Broekmans, J. Staal, and E. te Velde, "Computer assisted human follicle analysis for fertility prospects with 3d ultrasound," in *Information Processing in Medical Imaging*, ser. Lecture Notes in Computer Science, A. Kuba, M. Łamal, and A. Todd-Pokropek, Eds. Springer Berlin / Heidelberg, 1999, vol. 1613, pp. 56–69.

[28] B. Rahmatullah, I. Sarris, A. Papageorghiou, and J. A. Noble, "Quality control of fetal ultrasound images: Detection of abdomen anatomical landmarks using adaboost," in *Proceedings of the 8th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2011, March 30 - April 2, 2011, Chicago, Illinois, USA*. IEEE, 2011, pp. 6–9.

[29] S. Feng, S. Zhou, S. Good, and D. Comaniciu, "Automatic fetal face detection from ultrasound volumes via learning 3d and 2d information," in *Proc. CVPR*, Miami, FL, 20–25 Jun. 2009, pp. 2488–2495.

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

[31] R. Juang, E. McVeigh, B. Hoffmann, D. Yuh, and P. Burlina, "Automatic segmentation of the left-ventricular cavity and atrium in 3d ultrasound using graph cuts and the radial symmetry transform," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, 30 2011-april 2 2011, pp. 606 –609.

[32] Y. Zhan and D. Shen, "Automated segmentation of 3d us prostate images using statistical texture-based matching method," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003*, ser. Lecture Notes in Computer Science, R. Ellis and T. Peters, Eds. Springer Berlin / Heidelberg, 2003, vol. 2878, pp. 688–696.

[33] W. Lu, J. Tan, and R. Floyd, "Automated fetal head detection and measurement in ultrasound images by iterative randomized hough transform," *Ultrasound in Medicine & Biology*, vol. 31, no. 7, pp. 929 – 936, 2005.

[34] V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Trans. Med. Imag.*, vol. 16, no. 5, pp. 642–652, 1997.

[35] L. Zhang, S. Chen, C. T. Chin, T. Wang, and S. Li, "Intelligent scanning: Automated standard plane selection and biometric measurement of early gestational sac in routine ultrasound examination," *Medical Physics*, vol. 39, no. 8, pp. 5015–5027, 2012.

[36] 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. Imag.*, vol. 27, no. 9, pp. 1342–1355, Sept. 2008.

[37] Z. Tu, K. L. Narr, P. Dollr, I. Dinov, P. M. Thompson, and A. W. Toga, "Brain anatomical structure segmentation by hybrid discriminative/generative models," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 495–508, 2008.

[38] J. Anquez, E. Angelini, and I. Bloch, "Automatic segmentation of head structures on fetal MRI," in *Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. IEEE International Symposium on*, 28 2009-july 1 2009, pp. 109 –112.

[39] F. Rousseau, O. Glenn, B. Iordanova, C. Rodriguez-Carranza, D. Vigneron, J. Barkovich, and C. Studholme, "A novel approach to high resolution fetal brain MR imaging," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2005*, 2005, pp. 548–555.

[40] J. Zhang, S. Zhou, L. McMillan, and D. Comaniciu, "Joint real-time object detection and pose estimation using probabilistic boosting network," in *Proc. CVPR*, Minneapolis, MN, 18–23 Jun. 2007.

[41] J. S. Liu, R. Chen, and T. Logvinenko, "A theoretical framework for sequential importance sampling with resampling," in *Sequential Monte Carlo methods in practice*, A. Doucet, N. D. Freitas, and N. Gordon, Eds. Birkhäuser, 2001, pp. 225–242.

[42] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Medical Image Analysis*, vol. 17, no. 8, pp. 1293–1303, 2013.

[43] A. Torralba, K. Murphy, and W. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *Proc. CVPR*, vol. 2, 2004, pp. 762–769.

[44] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. CVPR*, Miami, FL, 20–25 Jun. 2009, pp. 1271–1278.

[45] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," in *Proc. ICCV*, Kyoto, Japan, 27 Sep–4 Oct 2009, pp. 229–236.

[46] S. Kumar and M. Hebert, "Discriminative random fields: a discriminative framework for contextual interaction in classification," in *Proc. ICCV*, vol. 2, 2003, pp. 1150–1157.

[47] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective," *International Journal of Computer Vision*, vol. 80, no. 1, pp. 3–15, Oct. 2008, 10.1007/s11263-008-0137-5.

[48] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. CVPR*, Anchorage, AK, 24–26 Jun. 2008, pp. 1–8.

[49] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *In NIPS*. MIT Press, 2004, pp. 1097–1104.

[50] N. J. Butko and J. R. Movellan, "Optimal scanning for faster object detection," in *Proc. CVPR*, Miami, FL, 20–25 Jun. 2009, pp. 2751–2758.

[51] Y. Zhan, X. Zhou, Z. Peng, and A. Krishnan, "Active scheduling of organ detection and segmentation in whole-body medical images," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2008*, 2008, pp. 313–321.

[52] D. Liu, S. Zhou, D. Bernhardt, and D. Comaniciu, "Search strategies for multiple landmark detection by submodular maximization," in *Proc. CVPR*, San Francisco, CA, 13–18 Jun. 2010, pp. 2831–2838.

[53] M. Sofka, K. Ralovich, N. Birkbeck, J. Zhang, and S. Zhou, "Integrated detection network (IDN) for pose and boundary estimation in medical images," in *Proceedings of the 8th International Symposium on Biomedical Imaging (ISBI 2011)*, Chicago, IL, 30 Mar – 2 Apr 2011.

[54] M. Sofka, K. Ralovich, J. Zhang, S. Zhou, and D. Comaniciu, "Progressive data transmission for anatomical landmark detection in a cloud," *Methods of Information in Medicine*, vol. 51, no. 3, pp. 268–278, 2012.

[55] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, ser. Monographs on Statistics and Applied Probability. Chapman and Hall, 1993, no. 57.

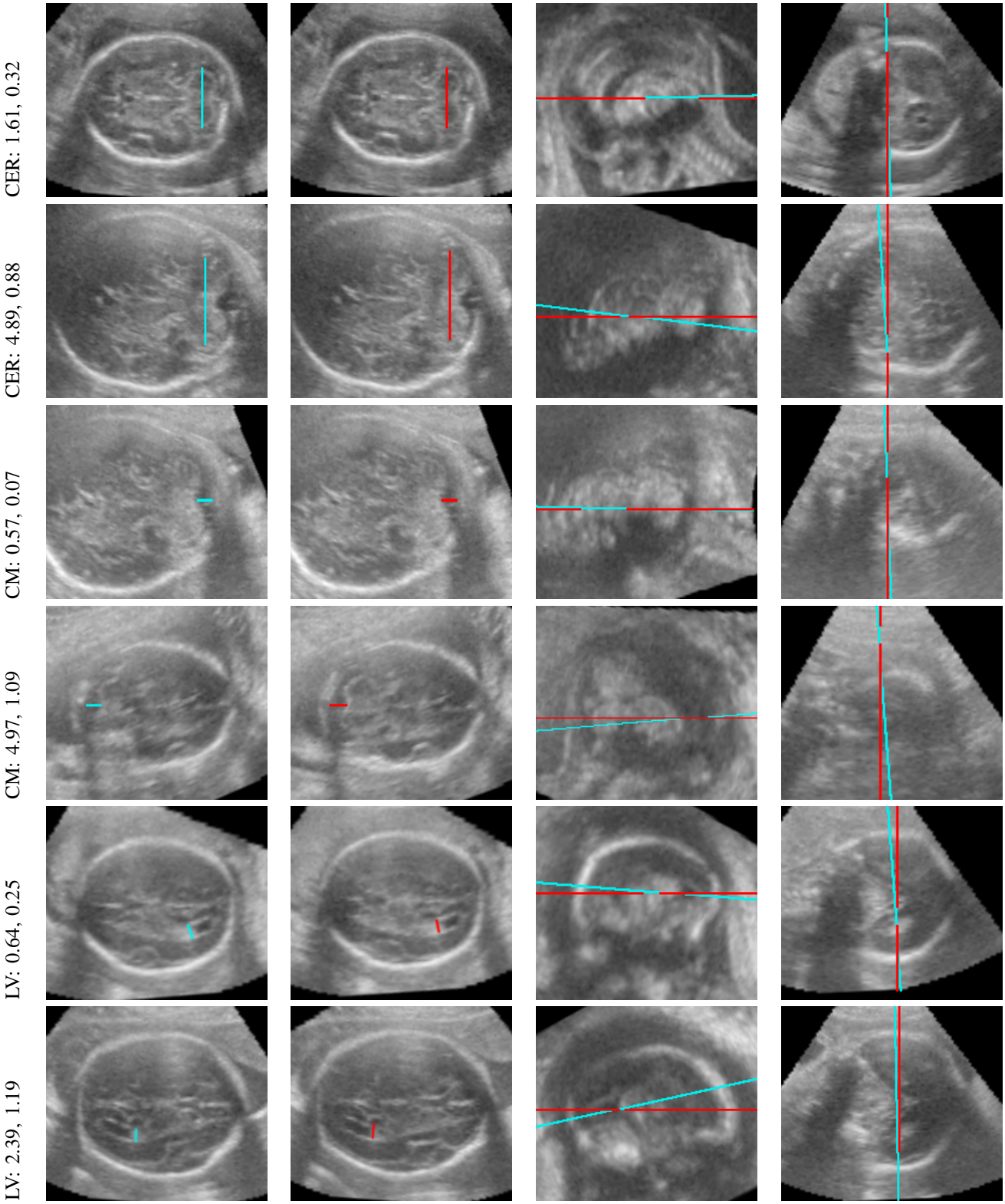

Fig. 8. The automatic measurement results (cyan) compared to ground truth (red) for Cerebellum (CER), Cisterna Magna (CM), and Lateral Ventricles (LV). The first row of each structure shows results with approximately *average* plane error and the second row shows results with *higher than average* plane error. Plane and length errors on the left are reported (in mm). Note, that the cases with higher error are still acceptable for clinical use. The last two columns show the agreement of the detection plane in the sagittal and coronal cross section.

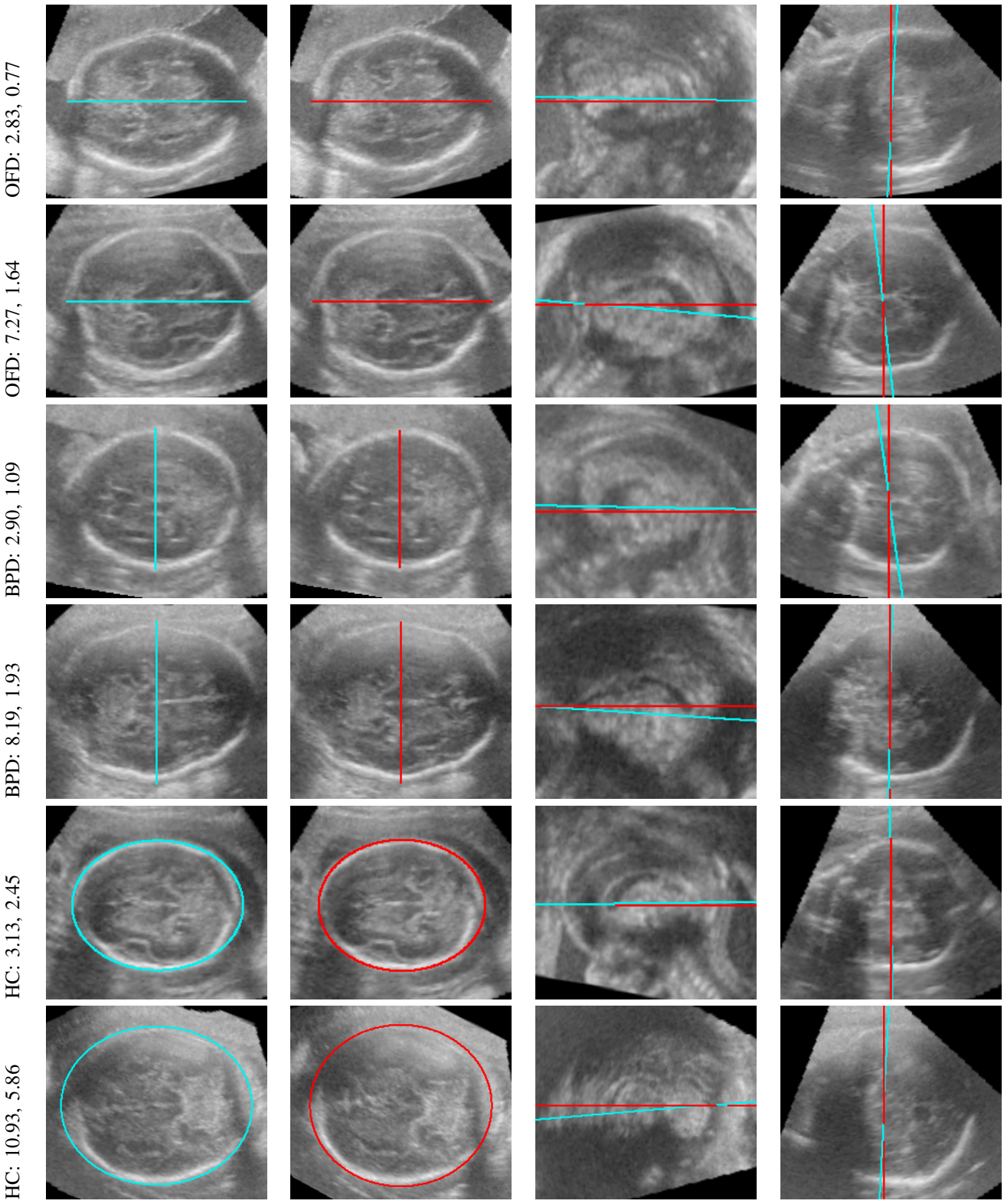

Fig. 9. The automatic measurement results (cyan) compared to ground truth (red) for Occipitofrontal Diameter (OFD), Biparietal Diameter (BPD), and Head Circumference (HC). Plane and length errors on the left are reported (in mm). See the text for details.

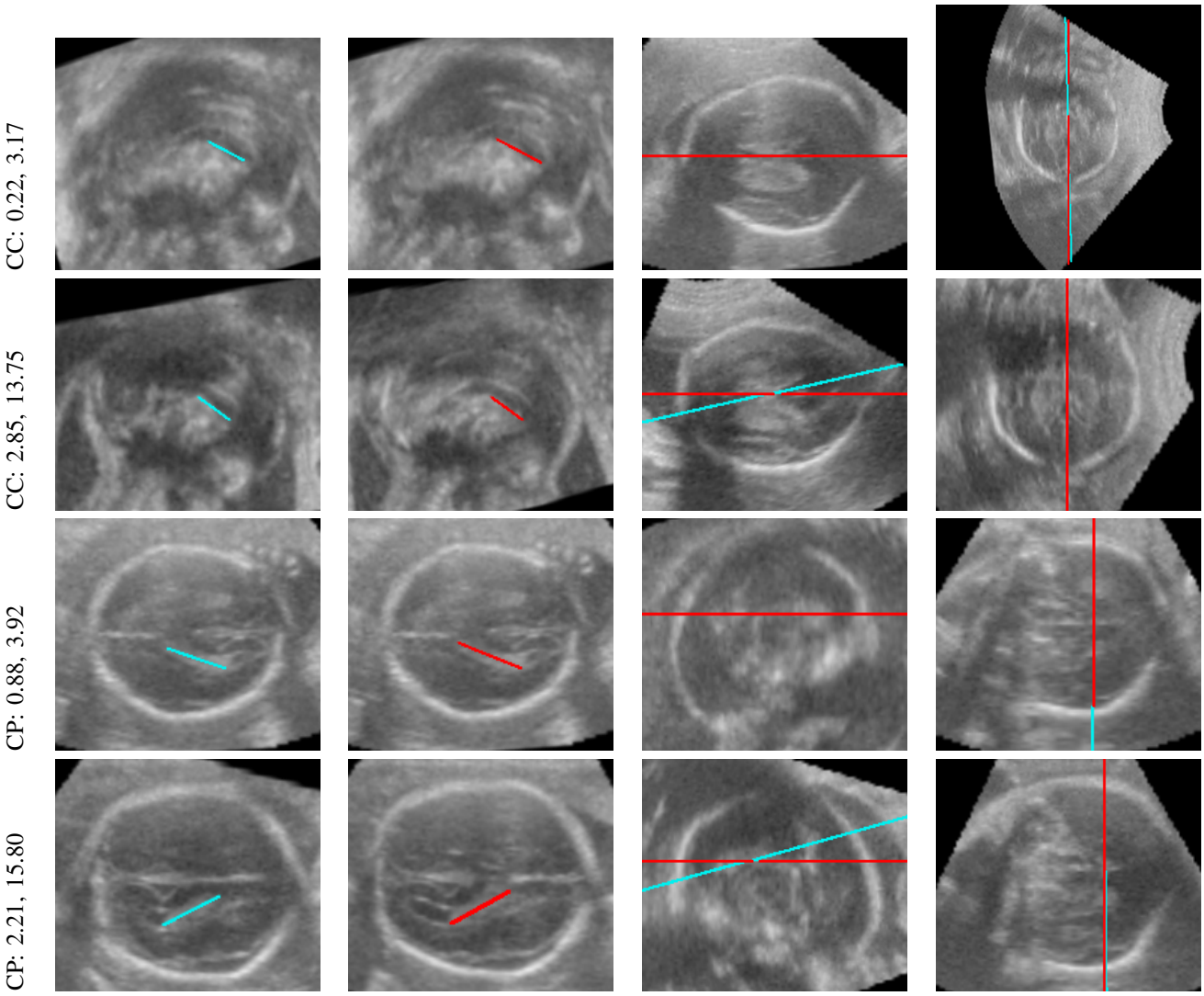

Fig. 10. The automatic detection results (cyan) compared to ground truth (red) for Corpus Callosum (CC) and Choroid Plexus (CP). The errors of detecting center and plane orientation are indicated (in mm). Annotation lines are shown for reference but they are not used clinically. See the text for details.

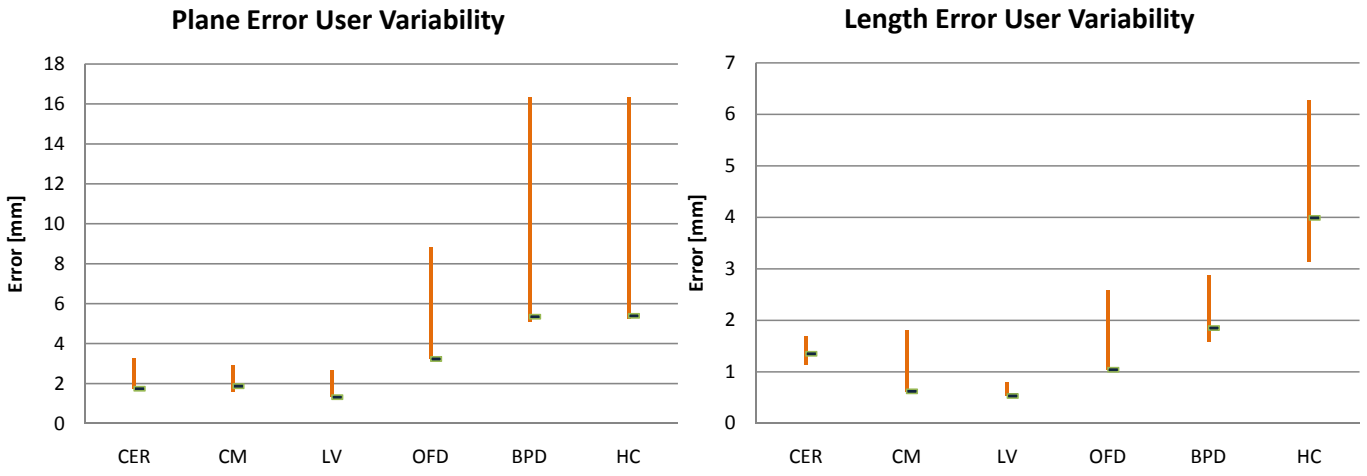

Fig. 11. The ranges of plane and length measurement errors (minimum and maximum) computed as an average over the 10 users and for each brain structure. The plots also show the deviation of each user w.r.t. initial Auto Fetal Head and Brain (AFHB) result with a green marker. The user variability can be quite large, especially for the longer measurements. The average changes to the results provided by AFHB system are small.