Deep Learning Based Spine Centerline Extraction in Fetal Ultrasound

Astrid Franz, Alexander Schmidt-Richberg, Eliza Orasanu, Cristian Lorenz

Philips GmbH Innovative Technologies, Röntgenstr. 24-26, D-22335 Hamburg
astrid.franz@philips.com

Abstract. Ultrasound is widely used for fetal screening. It allows for detecting abnormalities at an early gestational age, while being time and cost effective with no known adverse effects. Searching for optimal ultrasound planes for these investigations is a demanding and time-consuming task. Here we describe a method for automatically detecting the spine centerline in 3D fetal ultrasound images. We propose a two-stage approach combining deep learning and classic image processing techniques. First, we segment the spine using a deep learning approach. The resulting probability map is used as input for a tracing algorithm. The result is a sequence of points describing the spine centerline. This line can be used for measuring the spinal length and for generating view planes for the investigation of anomalies.

1 Introduction

Ultrasound is the modality of choice for fetal screening as it is able to show fetal anatomy in sufficient detail, while at the same being time and cost effective with no known adverse effects. Fetal screening allows for detecting abnormalities at an early gestational age, such that therapeutically suitable interventions can be planned and performed as required. Currently, the trend goes towards using 3D ultrasound, since a 3D image contains much more spatial information about the location of several organs with respect to each other and it allows for a variety of workflow optimizations.

One of the main fetal scans takes place in the second trimester between 18 and 22 weeks gestational age, when specific recommended standard measurements are determined, see for instance [1]. The localization of the abdominal cross-sectional plane with the corresponding measurement of the abdominal circumference is described in [2]. Another structure that is investigated is the fetal spine. First, the length of the spine provides an insight into the fetal growth. Second, a variety of spinal anomalies as spina bifida, meningocele, diastematomyelia, vertebral segmentation anomalies, sacral agenesis, spinal dysgenesis, spondylothoracic or spondylocostal dysplasia can be detected in 3D fetal ultrasound scans (see [3]). 3D ultrasound of the fetal spine has great diagnostic benefit, as it allows the operator to manipulate data in any plane after the completion of the exam.

State-of-the-art fetal ultrasound examination is based on a manual search for optimal view planes, which is a demanding and time-consuming task. Hence an automatic detection of the spine centerline would be of huge benefit, since it allows for an automatic generation of view planes which are well-suited for the investigation of spinal anomalies.

2 Materials and methods

Here we describe a method for automatically detecting the spine centerline in 3D fetal ultrasound images. We propose a two-stage approach combining deep learning and classic image processing techniques. First, we segment the spine using an F-NET approach, as described in [4]. The resulting probability map is used as input for a tracing algorithm, similar to the one described in [5], where this was used for rib centerline tracing in CT. The result is a sequence of points describing the spine centerline.

2.1 Data

Our data consists of 400 3D fetal ultrasound scans with a gestational age ranging from 15 weeks to 41 weeks (mean gestational age 26 weeks). The in-plane voxel size ranges from 0.12 mm to 0.51 mm (mean 0.25 mm) and the slice thickness from 0.30 mm to 0.93 mm (mean 0.51 mm). The data includes a wide range of typical ultrasound artifacts, such as shadows. These artifacts make it impossible to segment the spine directly, since some parts of the spine are extinguished. In all these datasets, the spinal canal is manually annotated by setting 5 to 10 landmarks, which can be done with reasonable effort. These landmarks are automatically connected by straight lines. Based on these centerline annotations, we generated label masks by dilating with a certain radius, given in number of voxels. Since this dilation radius is a crucial parameter in the whole spine segmentation pipeline, we varied this parameter between 0 and 50. To get a feeling of the thickness of the spine label map with respect to the dilation radius, Fig. 1(b) shows in blue as example a dilation radius 15. As can be seen, this radius nicely covers the spinal canal and parts of the surrounding bone. A dilation radius of 50 would cover the whole spine.

2.2 Preprocessing

As a preprocessing step, the images are isotropically resampled. Since the size of the fetus greatly varies with respect to the gestational age, we do not use a fixed voxel size, but choose a voxel size which is linearly growing with respect to the gestational age, starting with 0.4 mm at a gestational age of 14 weeks and ending at 1.6 mm at a gestational age of 42 weeks. These values were chosen with respect to fetal growth curves (see [6]). The spinal length in such resampled images averages at about 140 voxels, independent of the gestational age.

Table 1. Voxel size (vs), padding (p) and size of the receptive field (rf) of the F-NET at different resolution levels.

gestational age	resolution 1			resolution 2			resolution 3		
	vs	p	rf	vs	p	rf	vs	p	rf
14 weeks	0.4 mm	4	3.2 mm	0.8 mm	9	15.2 mm	1.6 mm	19	62.4 mm
42 weeks	1.6 mm	4	14.4 mm	3.2 mm	9	60.8 mm	6.4 mm	19	249.6 mm

2.3 Spine probability map generation

For spine detection, we first apply an F-NET as described in [4]. This is a fully convolutional network architecture that processes images at multiple scales and different fields of view in order to combine detailed voxel-level features with features that take large context into consideration, while requiring significantly less memory to store hidden layers than previously proposed U-NET architectures using shortcut connections (see [4]). The segmentation results were nearly independent of the F-NET parameters such as kernel size and number of feature extraction blocks, hence we used the standard settings of [4], namely kernel size 3 and number of feature extraction blocks 3. However, we reduced the number of resolution levels from the standard settings of 4 to 3, since the receptive field at resolution level 3 has a size of 25 cm at a gestational age of 42 weeks (see Tab. 1) and hence contains nearly the complete fetal torso. Thus the resolution level 4 would not add any more information. In this way, we could enlarge the patch size, i.e. the number of voxels the network can process in parallel, and hence speed up the computation.

2.4 Centerline tracing

The result of the previous step is a probability map describing the probability of the class spine. This probability map is used as input for an iterative tracing algorithm, as described in [5]. During the tracing, a probability weighted covariance matrix $\Sigma_i \in \mathbb{R}^{3\times3}$ of the voxel coordinates in a spherical region around the point of interest c_i is evaluated to estimate the tracing direction. Furthermore, the same sphere is used to calculate a probability-weighted center of mass. The tracing starts off with the point $c_0 \in \mathbb{R}^3$ having largest probability. If other points share this probability, the first one found is taken. First at c_0 the radius of the averaging sphere is chosen such that the fraction of voxels $F_{p>0.5}$ with probability $p > 0.5$ is between predefined limits $F_{\min} < F_{p>0.5} < F_{\max}$, where limits have been chosen empirically to be $F_{\min} = 0.02$ and $F_{\max} = 0.06$. Then, keeping the chosen radius of the averaging sphere, points are iteratively added to the centerline by repeating the following steps for $i = 0, 1, 2, \ldots$:

1. Calculate the probability weighted covariance matrix at c_i. The eigenvector $t_i \in \mathbb{R}^3$ corresponding to the largest eigenvalue of Σ_i is used as estimation of the tangent direction.
2. Starting from c_i, move in tangential direction t_i until a probability value < 0.5 or a maximal moving distance $d_{\max} = 15$ voxels is reached. If before

reaching $d_{\max}$, the probability drops below < 0.5, a look-ahead mechanism tries to bridge the gap in walking direction up to a maximal distance $d_{\text{gap}} = 10$ voxels.
3. Calculate the weighted mean vector $c_{i+1} \in \mathbb{R}^3$ in a spherical region around the current position and add this point to the list of spine centerline points.

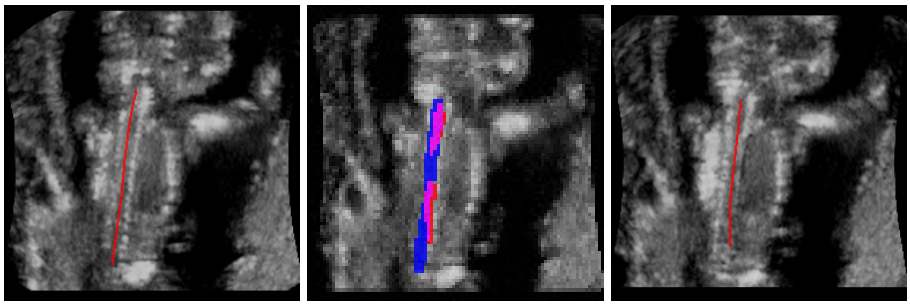
This pipeline is iterated until the moving distance falls below a predefined threshold. Tracing from the initial point c_0 is performed in both possible directions, and the results are concatenated.

2.5 Evaluation metrics

For evaluating the results, we used three different evaluation metrics.

Dice coefficient We used the Dice coefficient for evaluating the F-NET segmentation, described in Sec. 2.3. The Dice coefficient measures the overlapping region of the thickened ground truth annotation and the probability map generated by the F-NET. It is a number between 0 and 1, where 0 means no overlap at all and 1 means full overlap.

False negative rate The false negative rate (FNR) describes the fraction of the ground truth line which is not contained in the resulting traced spine centerline. We used a pre-defined outlier distance of 10 mm to define which points are found, i.e. a point of the ground truth annotation is considered to be found if there is a point in the traced spine centerline with a distance of at most 10 mm. The FNR is given in % throughout this paper. For computing the FNR, both the ground truth line and the detected centerline are approximated by 100 tightly-spaced points, and for each of these 100 ground truth points it is checked if a detected centerline point is in the range of the outlier distance.



(a) (b) (c)

Fig. 1. Example of a test case: (a) ground truth annotation of the spine centerline, (b) thickened ground truth centerline (blue) and FNET segmentation result (red), (c) traced centerline resulting from the FNET probability map. The images (a) and (c) are a flattened view following the annotated line and are therefore not a slice of the original dataset, whereas (b) is an image slice containing most of the thickened ground truth centerline.

Mean distance For each of the 100 tightly-spaced points of the detected spine centerline, which is not a false negative as defined before, the distance to the ground truth line is computed. This distance is averaged yielding a mean distance of the detected spine centerline to the ground truth annotation.

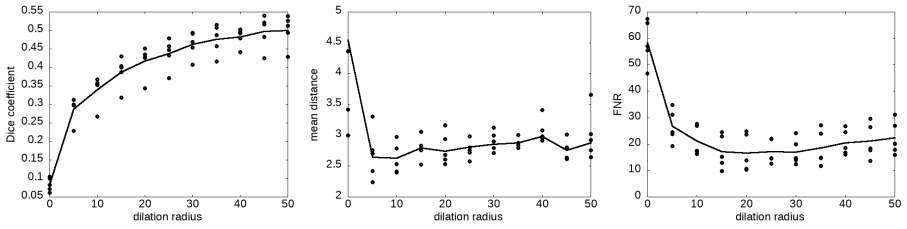
2.6 Experiment set-up

For evaluation, we split our dataset in five folds of 80 images each to perform a five-fold cross validation: we used four folds, i.e. 320 datasets, for network training, applied the resulting segmentation network to the remaining fold of 80 test datasets, and subsequently traced the centerline of the resulting probability maps for the test datasets. One example case is shown in Fig. 1.

3 Results

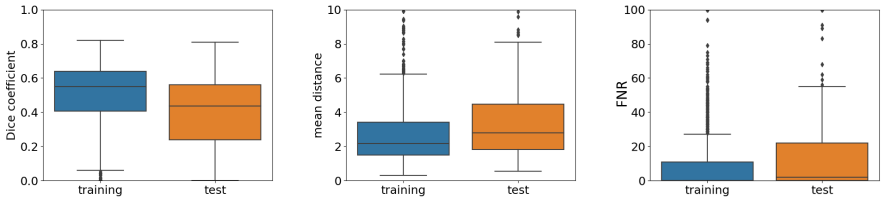
The evaluation measures (see Sec. 2.5) for a variety of dilation radii used for constructing ground truth label masks out of the annotated spine centerlines are shown in Fig. 2.

Fig. 3 shows the distribution of the Dice coefficient, the mean distance and the FNR for all our 400 cases for a dilation radius of 15 as a box-whisker-plot, both for training and test.



(a) (b) (c)

Fig. 2. (a) Dice coefficient, (b) mean distance and (c) FNR as a function of the dilation radius. The points indicate the values for the five folds, the line is the mean over all folds.



(a) (b) (c)

Fig. 3. Box-whisker-plots for the (a) Dice coefficient, (b) mean distance and (c) FNR for a dilation radius of 15.

4 Discussion

As can be seen in Fig. 2(a), the Dice coefficient is increasing with growing dilation radius. This is to be expected since the more voxels a structure has, the less a voxel mismatch will count towards the overlap. For elongated structures with fewer voxels, this measure is not well-suited for describing the segmentation accuracy: As soon as the detected centerline is off by a few millimeters, the Dice coefficient drops drastically. In the example of Fig. 1, the Dice coefficient is 0.539, although the spine centerline is correctly detected: The FNR is 0 meaning that all spine centerline points are detected within the outlier distance of 10 mm, and the mean distance of the detected centerline to the ground truth line in this case is 1.014 mm. Hence for the application of detecting the spine centerline, the FNR and the mean distance are better suited for evaluating the result. These two measures show a drastic decrease when the dilation radius goes from 0 to 5, and then stay nearly constant, where the FNR shows a minimum at a dilation radius of approx. 15 (see Fig. 2(c)). Hence a dilation radius of 15 seems to be reasonable for this application.

Fig. 3 concentrates on a dilation radius of 15. Clearly the Dice coefficient is larger for training (see Fig. 3(a)), while the mean distance and the FNR values are smaller (see Fig. 3(b,c)), but these effects are not very pronounced, indicating only a slight over-fitting effect. Half of the test cases have a FNR below 3, meaning that in half of the test cases at least 97% of the annotated ground truth spine line is found by our algorithm.

Hence our presented fully automatic two-stage approach for detecting the spine centerline in 3D fetal ultrasound scans yields promising results. The resulting spine centerline can be used for selecting dedicated view planes. They can be presented to the physician for investigating the spine, for instance for measuring the length or detecting anomalies. An estimation of the spinal length can even automatically be proposed by calculating the length of the detected centerline. Furthermore, a flex view could be generated showing the spine and the outgoing ribs in a straightened view for further investigation.

References

1. Papageorghiou AT, (et al). International standards for fetal growth based on serial ultrasound measurements: The Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. The Lancet. 2014;384:869–879.
2. Lorenz C, (et al). Automated abdominal plane and circumference estimation in 3D US for fetal screening. Procs SPIE. 2018;10574:105740I.
3. Upasani V, (et al). Prenatal diagnosis and assessment of congenital spinal anomalies: Review for prenatal counseling. World J Orthop. 2016;7:406–417.
4. Brosch T, Saalbach A. Foveal fully convolutional nets for multi-organ segmentation. Procs SPIE. 2018;10574:105740U.
5. Lenga M, (et al). Deep learning based rib centerline extraction and labeling. Lecture Notes in Computer Science. 2018;11404:99–113.
6. Ulm M, (et al). Ultrasound evaluation of fetal spine length between 14 and 24 weeks of gestation. Prenatal Diagnosis. 1999;19:637–641.