Computer Methods and Programs in Biomedicine (2020)

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.journals.elsevier.com/computer-methods-and-programs-in-biomedicine

A regression framework to head-circumference delineation from US fetal images

Maria Chiara Fiorentino[a], Sara Moccia[a,b,*], Morris Capparuccini[a], Sara Giamberini[a], Emanuele Frontoni[a]

[a] *Department of Information Engineering, Universita Politecnica delle Marche,Via Brecce Bianche, 12, 60131 Ancona, Italy*
[b] *Department of Advanced Robotics, Istituto Italiano di Tecnologia, Via Morego, 30, 16163 Genova, Italy*

ARTICLE INFO

Article history:

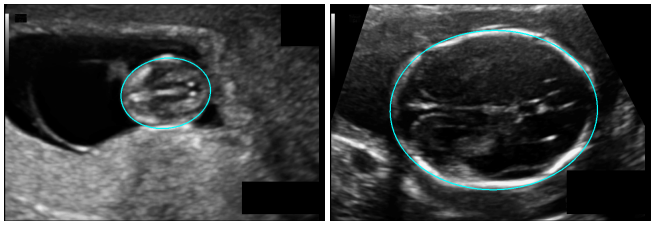
Keywords: Fetal ultrasounds, Head circumference delineation, Regression networks, Convolutional neural networks

ABSTRACT

Background and Objectives Measuring head-circumference (HC) length from ultrasound (US) images is a crucial clinical task to assess fetus growth. To lower intra- and inter-operator variability in HC length measuring, several computer-assisted solutions have been proposed in the years. Recently, a large number of deep-learning approaches is addressing the problem of HC delineation through the segmentation of the whole fetal head via convolutional neural networks (CNNs). Since the task is a edge-delineation problem, we propose a different strategy based on regression CNNs. **Methods** The proposed framework consists of a region-proposal CNN for head localization and centering, and a regression CNN for accurately delineate the HC. The first CNN is trained exploiting transfer learning, while we propose a training strategy for the regression CNN based on distance fields. **Results** The framework was tested on the *HC18 Challenge* dataset, which consists of 999 training and 335 testing images. A mean absolute difference of 1.90 (± 1.76) mm and a Dice similarity coefficient of 97.75 (± 1.32) % were achieved, overcoming approaches in the literature. **Conclusions** The experimental results showed the effectiveness of the proposed framework, proving its potential in supporting clinicians during the clinical practice.

1. Introduction

Fetal biometrics have a strong diagnostic and prognostic role in the evaluation of fetal growth [1]. Fetal biometrics are commonly measured in the clinical practice from obstetric ultrasound (US), which is a non-invasive, low-cost imaging modality that allows real-time image acquisition. Among fetal biometrics, head-circumference (HC) length is often measured by clinicians to estimate gestational age and fetal weight, especially when abnormal head growth is suspected [2]. In the clinical practice, HC-length measurement is performed manually, either overlaying an ellipse on the fetal skull or identifying landmarks on the skull that delimit the head main axes.



(a) First trimester (b) Second trimester

Fig. 1: Samples of ultrasound images acquired at the (a) first and (b) second trimester. The head-circumference annotations in shown in light blue.

*Corresponding author: Tel.: +39 071 2204825
e-mail: s.moccia@staff.univpm.it (Sara Moccia)

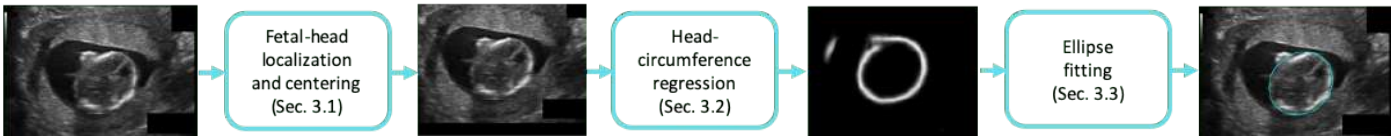


Fig. 2: Workflow of the proposed framework for head-circumference delineation.

The manual delineation, however, poses issues related to both measurement reproducibility and time consumption [3].

A possible solution to attenuate these issues would be to develop an algorithm for automatic HC delineation. However, this is not a trivial task: challenges that have to be addressed include presence of shadows in the images and signal dropout and speckling, which cause fake or missing edges [4]. As a further challenge, the HC only covers a small portion of the image, which moreover varies among the trimesters (Fig. 1).

Recently, researches in closer fields (e.g., [5], [6], [7], [8]) have pointed out the benefits of modeling edge-delineation problems, such as HC delineation, as heatmap regression tasks. Following such modeling, a convolutional neural network (CNN) is trained to directly regress a distance-field from the edge of interest.

Thus, guided by the research hypothesis that regression CNNs may allow accurate HC delineation, the contributions of this paper can be summarized as follows:

1. A new framework for HC delineation, which consists of a region-proposal CNN for head location and centering, and a regression CNN for accurately delineate the HC
2. A new strategy to train the regression CNN that is based on distance fields

2. Related work

Considering the relevance of HC measurement in the clinical practice, several automatic and semi-automatic algorithms have been proposed in the years. First approaches were mainly model-based and relied on randomized Hough transform ([9], [10]), circular shortest path [11], active contours ([12], [13]) and Chan-Vese level set [14].

More recently, machine learning has been investigated to tackle the peculiar challenges of US images, as listed in Sec. 1. A large variety of handcrafted features has been studied. Examples include distance-[15] and intensity-based features [16], Haar descriptors ([17], [18], [19]) and textons [20].

To foster researches following the learning paradigm, one challenge was realized in 2012, with the release of a dataset of 90 US images. More recently, in 2018, the *HC18 Grand Challenge* was organized, with the release of a dataset of 1334 US images (divided in 999 training and 335 test images). Such dataset size allows the development of more advanced solutions.

In fact, the handcrafted-based approaches have been overcame by deep learning (DL) [21, 22, 23], which directly extracts features from raw data, avoiding their explicit mathematical formulation, but requires larger training datasets [24]. The work in [25] proposes an active-contour model guided by external forces that are derived with a CNN to segment the fetal head. The HC is then identified by ellipse fitting with the direct least squares fitting of ellipses method proposed in [26].

Nonetheless, the formulation of the external forces influence the performance of the method, directly.

Region proposal networks, such as Yolo, are used in [27] to limit the image area to be processed for HC search. Hough transform and a filter-based approach relying on the difference of Gaussians are then used to delineate the HC, posing issues for parameter tuning (both for the Hough transform and the filtering).

A large number of researchers have proposed CNN models for segmenting the whole fetal head, as a preliminary step for HC delineation. [28] propose a CNN inspired by LinkNet [29], where the LinkNet encoder path is used to perform feature extraction but also to obtain, thought fully connected neurons, the HC main axes, center and angle. The main disadvantage is that directly regressing the HC pixel position is empirically too localized, making this type of approach particularly difficult. Sobhanina et al. present also another approach [30] based on a multiscale mini-LinkNet network to perform fetal-head segmentation, achieving promising results due to the multiscale feature representation. In [31], polar transformation followed by UNet is applied to segment head boundary. A region-proposal network is then used as post processing to remove wrongly segmented pixels. Even if with encouraging results, the approach was tested on 102 images only.

A similar approach is proposed by [32], where two UNet are exploited in sequence for rough skull segmentation, and segmentation refinement, respectively, from 3D fetal US.

The work in [33] develops two probabilistic CNN methods: Monte Carlo Dropout during inference and a Probabilistic UNet. An ensemble of the generated segmentation masks is used to reject acquired images that produce sub-optimal HC measurements. These methods could be particularly useful in the clinical practice since multiple plausible semantic segmentation hypotheses are provided to the clinicians, which can choose the best option. In [34], a combined fetal-head localisation and fetal-head segmentation approach based on Mask R-CNN is proposed. In [31], [15] and [34] the HC is then identified by least square ellipse fitting method.

Most of these approaches mainly reduces the problem of HC delineation to that of segmenting the whole fetal head via CNNs. A different strategy could be to directly regress HC pixel position. However, this is a highly non-linear problem, which poses challenges to training convergence [35]. Regressing a distance field from the HC could attenuate this issue. This concept has recently been proved to be successful in closer fields ([5], [6], [7]). Guided by these recent considerations, in this work we will investigate the concept of distance-field regression to accurately delineate HC from US images.

Table 1: Regression network architecture with Conv2D = convolution, MaxPool2D = max pooling, Upsamp = upsampling, Concat = concatenation.

	Name	Feature maps (inputs)	Feature maps (output)
Contraction section	Conv2D	192x240x1	192x240x64
	Conv2D	192x240x64	192x240x64
	MaxPool2D	192x240x64	96x120x64
	Conv2D	96x120x64	96x120x128
	Conv2D	96x120x128	96x120x128
	MaxPool2D	96x120x128	48x60x128
	Conv2D	48x60x128	48x60x256
	Conv2D	48x60x256	48x60x256
	MaxPool2D	48x60x256	24x30x256
	Conv2D	24x30x256	24x30x512
	Conv2D	24x30x512	24x30x512
	Dropout	24x30x512	24x30x512
	MaxPool2D	24x30x512	12x15x512
Bottleneck	Conv2D	12x15x512	12x15x1024
	Conv2D	12x15x1024	12x15x1024
	Dropout	12x15x1024	12x15x1024
	Upconv	12x15x1024	24x30x1024
Expansion section	Conv2D	24x30x1024	24x30x512
	Concat	24x30x512	24x30x1024
	Conv2D	24x30x1024	24x30x512
	Conv2D	24x30x512	24x30x512
	Upconv	24x30x512	48x60x512
	Conv2D	48x60x512	48x60x256
	Concat	48x60x256	48x60x512
	Conv2D	48x60x512	48x60x256
	Conv2D	48x60x256	48x60x256
	Upconv	48x60x256	96x120x256
	Conv2D	96x120x256	96x120x128
	Concat	96x120x128	96x120x256
	Conv2D	96x120x256	96x120x128
	Conv2D	96x120x128	96x120x128
	Upconv	96x120x128	192x240x128
	Conv2D	192x240x128	192x240x64
	Concat	192x240x64	192x240x128
	Conv2D	192x240x128	192x240x64
	Conv2D	192x240x64	192x240x64
	Conv2D	192x240x64	192x240x2
	Conv2D	192x240x2	192x240x1

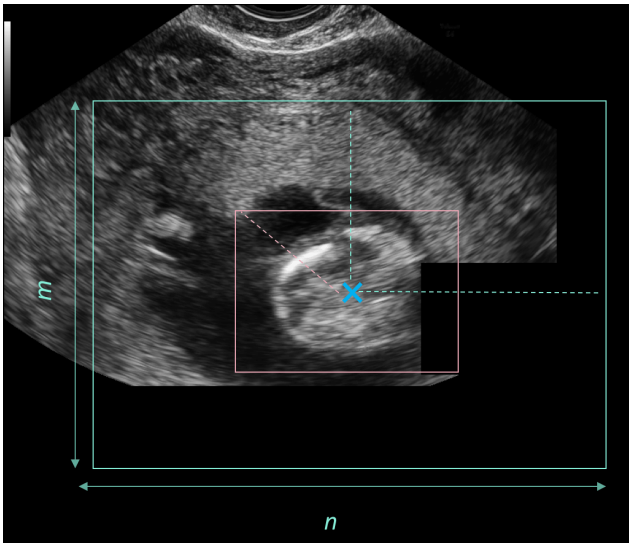


Fig. 3: Fetal-head localization and centering: from the predicted bounding box (pink rectangle), the coordinates of the head-circumference (HC) centre are estimated. The centered image (light-blue rectangle) is obtained by centering the original image with respect to the HC center. The centered image has dimension *nxm* (with *n* and *m* equal to 705 and 545, respectively) to match the biggest bounding box found in the training set. Zero padding is added when the HC is too close to image borders.

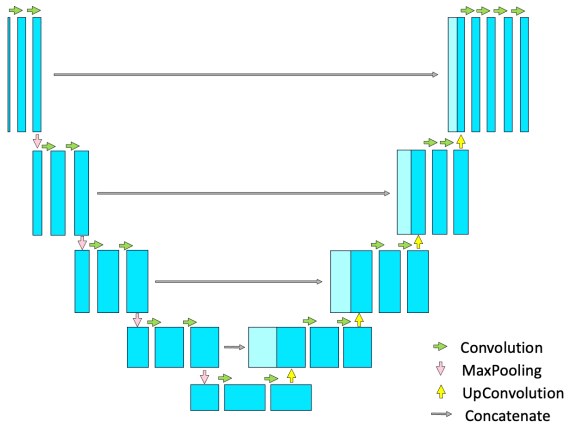
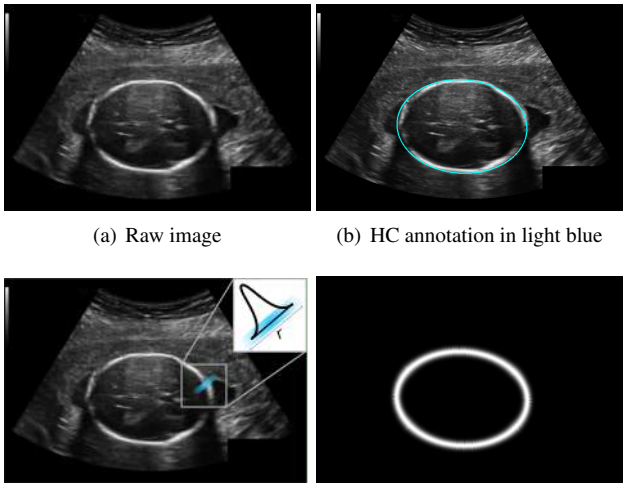


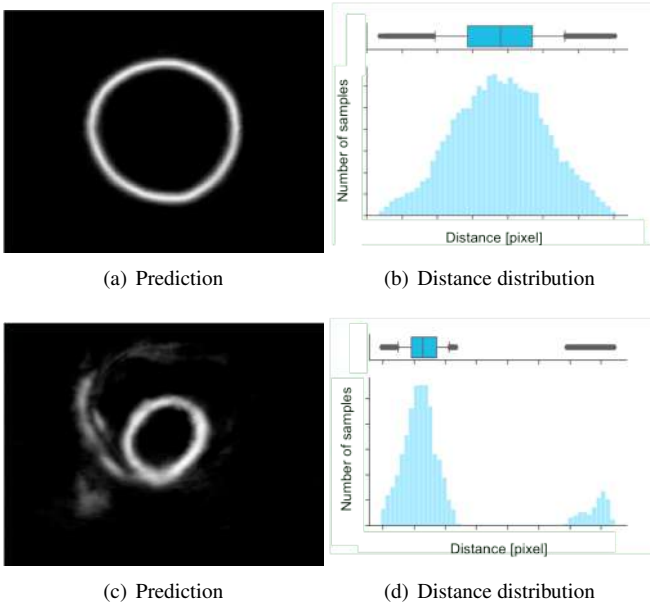
Fig. 4: UNet architecture. The architecture includes four contraction blocks along with a bottleneck section and four expansion blocks, as described in Sec. 3.2.



(a) Raw image (b) HC annotation in light blue

(c) Gaussian profile. (d) Training mask.

Fig. 5: Ground-truth generation for the regression network. The raw image in (a) is annotated by drawing the ellipse representing the head circumference (HC) showed in light blue in (b). The distance field is then built following a Gaussian profile (thickness r pixels), as shown in (c), to obtain the training mask showed in (d).



(a) Prediction (b) Distance distribution

(c) Prediction (d) Distance distribution

Fig. 6: In (a) and (c), the regression output is shown for two test images. The corresponding distance distributions are shown in (b) and (d), respectively, with the corresponding boxplot and outliers on top.

To the best of our knowledge, this is the first attempt to investigate regression CNN for automatic HC delineation from US images. We developed and tested the proposed approach using the *HC18 Grand Challenge*[1] dataset.

3. Methods

Figure 2 shows an overview of the workflow of the proposed framework. We exploit two consecutive CNNs for localizing and centering the fetal head (Sec. 3.1) and regressing the HC pixel position via distance fields (Sec. 3.2). The last step (Sec. 3.3) consists of fitting the output of the regression CNN with an ellipse, to mimic what clinicians do in the actual clinical practice. From the fitted ellipse, the HC length is computed.

3.1. Fetal-head localization and centering

As a preliminary step, inspired by [27], [31] and [34], we use an object-proposal localization CNN (i.e., the tiny-YOLOv2 [36]) for localizing and centering the fetal head. This way, the regression network is relieved of learning the position of the head.

The tiny-YOLOv2 is composed of 8 blocks, each consisting of a convolutional layer (kernel size = 3x3, and stride = 1). Each convolution is followed by batch normalization, and the resulting output is activated with a LeakyReLU (with a constant multiplier, α, equal to 0.1 to control the slope of the activation function for negative values). Max pooling (size = 2x2) is applied after each LeakyReLU activation in the first 4 blocks, with the goal to reduce the number of parameters to learn during training.

The tiny-YOLOv2 is here trained exploiting transfer learning and fine tuning. Specifically, we use, as initial weights, the weights obtained by training tiny-YOLOv2 on natural images. The weights are available online[2]. For fine-tuning purpose, the weights and bias of the last 4 convolutional layers are randomly re-initialized using a normal distribution. The adaptive moment estimation (Adam) is used to minimize the loss function proposed in the original YOLOv2 paper [36]. As shown in Fig. 3, we use the bounding box provided by the tiny-YOLOv2 to center the image before feeding the regression network. Keeping the fetal head centered, we select an image region with a size ($m \times n$) equal to the that of the largest bounding box found in the training set. When the region falls outside the image area (e.g., when the HC is close to image borders), zero padding is performed. This procedure is done to avoid changing the HC aspect ratio.

3.2. Head-circumference regression

The proposed strategy to train the regression CNN relies on distance fields. The rationale behind using distance field is to smooth the HC line as to facilitate the network task with respect to directly regressing the HC line.

The HC ground-truth masks of the *HC18 Grand Challenge* dataset are ~2-pixels wide. Hence, to build our distance field, the masks are skeletonized first. Then, as shown in Fig. 5, we consider a region consisting of all pixels that lie in the rectangular region with thickness r pixels, centrally aligned with each of the pixel of the HC, and perpendicular to the tangent of the HC. Inspired by the work in [5], which uses a regression network to estimate surgical-robot joint position, we used the original UNet architecture [37] modifying the output layer to accomplish a regression task. Choosing UNet can be twice as beneficial: in addition to showing astounding results in regards to medical-image segmentation [38], work in the literature for fetal-head segmentation (as described in Sec. 2) used U-shaped network, making the comparison with our regression architecture straightforward. The regression architecture, shown in Fig. 4 and described in Table 1, consists of a contraction section, a bottleneck and expansion section. The contraction section includes 4 blocks, each with two convolution layers (kernel size = 3x3 with no padding) followed by a ReLU and a 2x2 max pooling (stride = 2). The last block adds a Dropout layer (dropout rate = 0.5) between the convolutional and max-pooling layers. The number of channels is doubled at each block to incrementally learn more complex features, starting from 64 and reaching 512 channels.

The bottleneck connects the contraction section with the expansion section using two convolutional layers (kernel size = 3x3) followed by dropout (with a rate = 0.5) and one up convolutional layer (kernel size = 2x2). The number of channel doubles from 512 to 1024.

Similarly to the contraction section, the expansion section is made of 4 blocks. Each block consists of two convolutional layers (kernel size = 3x3) followed by a ReLU activation function and an upsampling layer (kernel size = 2x2) that halves

[1] http://doi.org/10.5281/zenodo.1322001

[2] https://github.com/experiencor/keras-yolo2.git

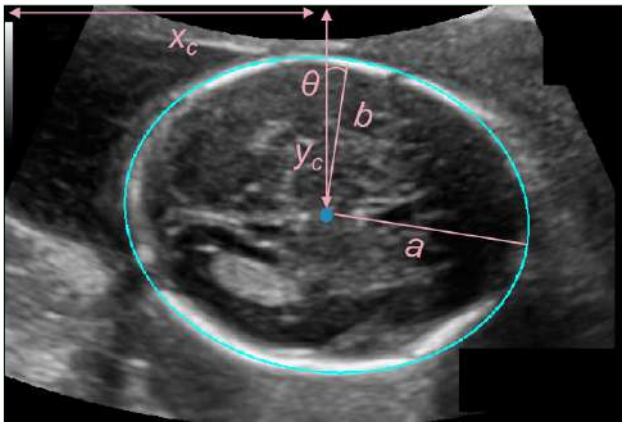


Fig. 7: Head-circumference (HC) parameters obtained from the fitted ellipse: (x_c, y_c) = HC center coordinates, a = semi-major axis length, b = semi-minor axis length, and θ = angle of orientation.

the number of feature channels. The input of each contraction section is concatenated with the correspondingly cropped feature map from the contracting section, as to recover the feature lost due to the downsampling in the contraction path. The last block is composed of four convolutional layers, three of which are followed by a ReLU activation function. As opposed to the UNet proposed by Ronneberger [37] in which a softmax activation function is used, the last convolutional layer is activated by the hyperbolic tangent (tanh).

The regression network is trained using the stochastic gradient descent with momentum as optimizer, minimizing the Mean Squared Error (MSE) as loss function:

$$MSE(M, \hat{M}) = \frac{1}{N}\sum_{i=0}^{N}(M_i - \hat{M}_i)^2 \quad (1)$$

where M_i and $\hat{M}_i$ are the regression mask and predicted outcome from the regression network for the i-th training image, respectively, and N is number of training samples.

3.3. Ellipse fitting

Before performing ellipse fitting, we threshold the output of the regression network. We then compute the distance between each point on the thresholded output and the center of the image, which matches the HC center, resulting in a distance distribution. As shown in Fig. 6, we identify (and remove) the outliers of the distance distribution as those values outside the range $[Q1 - O * IQR, Q3 + O * IQR]$ [39], where Q_1 and Q_3 are the first and third quartile, respectively, IQR is the interquartile range, and O is a constant value that is set according to the validation set (Sec. 4.2).

The resulting image is fitted with an ellipse using a geometric distance based method (i.e., ElliFit) [40], which is unconstrained, non-iterative and computationally inexpensive. From the fitted ellipse, the five parameters (shown in Fig. 7) that univocally identify it (i.e., semi-major axis length (a), semi-minor axis length (b), angle of orientation (θ), and center (x_c, y_c)) are obtained.

4. Experimental setup

4.1. Dataset

The dataset used to evaluate the proposed methodology was released in the context of the *HC18 Grand Challenge*[3]. The dataset consists of a training set of 999 and a test set of 335 US images from 551 women, acquired at the beginning of the first, second and third trimester of pregnancy. The images were collected at the Department of Obstetrics of the Radbound University Medical Center, Nijmegen, Netherlands, using both the Voluson E8 and the Voluson 730 (General Electric, Austria) [19]. All data were acquired by competent sonographers and anonymized as state in the Declaration of Helsinki. The local ethics committee (CMO Arnhem-Nijmegen) approved the collection and use of the data for research purposes.

Each image had a size of 800x540 pixels, with a pixel size ranging from 0.052 to 0.326 mm, due to sonoprapher adjustments to face the different size of fetuses.

For each image, the sonographer manually annotated the HC by drawing an ellipse that best fits the skull section. In this work, we kept 301 images (out of the 999 training images) as validation set.

The most peculiar challenges of the testing set are shown in Fig. 8. Challenges included different position of the head in the image, as well as varying dimension of the fetal head among the gestational trimesters.

Hence, the images of the first trimester showed a tiny head with skull edges not always visible. Reverberations and shadows were also be present, which resulted in poor fetal head contrast with respect to the background.

4.2. Parameter settings

To train the tiny-YOLOv2, the COCO challenge annotation format was followed. Starting from the HC annotation, we generated the bounding box that bounded the HC. Prior feeding tiny-YOLOv2, the images were resized to 800x800 pixels to match the dimension of the pretrained model. The tiny-YOLOv2 was fine tuned using Adam for 100 epochs, with an initial learning rate of 0.0001 and batch size equal to 16. Prior fine tuning, offline data augmentation (with horizontal flipping and rotation) was implemented to augment the number of images from the first semester.

The masks used for training the regression network were obtained with an r empirically set to 100 pixels and a Gaussian standard deviation of $r/2$. This allowed to fully cover the head-skull section at each HC point. Before feeding the regression network, we performed the preprocessing explained in Sec. 3.1, with m and n equal to 545 and 705 pixels, respectively.

The images were zero padded (from 182x235 to 192x240 pixels) to match the dimension required by the regression network. Also in this case, padding was preferred over reshaping for preserving the fetal-head aspect ratio.

The regression network was trained with mini-batch stochastic gradient descent, with an initial learning rate of 0.001 and

[3] https://hc18.grand-challenge.org/

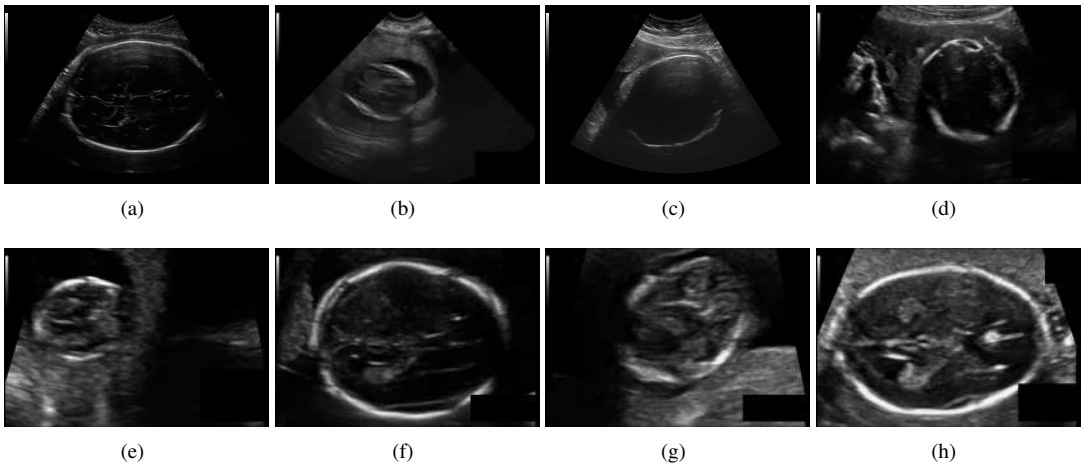


Fig. 8: Challenges in the testing datasets include different fetus head dimensions (a, b, h) and position (d, f) in the image, partially visible head skull (c), presence of noise and US image artifacts (e, g).

Table 2: Metric comparison for the ablation study. The mean is reported for each metric, with standard deviation in brackets.

	Absolute difference (AD) [mm]	Difference (DF) [mm]	Dice similarity coefficient (DSC) [%]	Hausdorff difference (HD) [mm]
Shallow regression	11.11 (± 38.25)	2.74 (± 39.74)	93.14 (± 12.23)	3.82 (± 6.04)
Proposed-short	5.67 (± 30.28)	3.62 (± 30.60)	95.58 (± 9.32)	2.53 (± 4.59)
Proposed-middle	2.08 (± 2.08)	-0.13 (± 2.04)	97.62 (± 4.17)	1.38 (± 1.19)
Regression only	5.79 (± 16.58)	5.27 (± 16.75)	95.33 (± 9.51)	2.82 (± 5.47)
Proposed without excluding outliers	2.33 (± 3.36)	1.46 (± 4.10)	97.30 (± 3.21)	1.51 (± 3.68)
Proposed	**1.90 (± 1.77)**	**0.21 (± 2.59)**	**97.76 (± 1.32)**	**1.32 (± 0.73)**

Table 3: Metric comparison for methods in the state of the art using the same test set. The mean is reported for each metric, with standard deviation in brackets.

	Absolute difference (AD) [mm]	Difference (DF) [mm]	Dice similarity coefficient (DSC) [%]	Hausdorff difference (HD) [mm]
[19]	2.80 (± 3.30)	0.60 (± 4.30)	97.00 (± 2.80)	2.00 (± 1.60)
[28]	2.12 (± 1.87)	1.13 (± 2.69)	96.84 (± 2.89)	1.72 (± 1.39)
[30]	2.22 (N.A.)	1.19 (N.A.)	92.46 (N.A.)	3.40 (N.A.)
[25]	2.45 (± 2.55)	-1.05 (± 3.38)	95.49 (± 4.11)	2.44 (± 1.96)
[34]	2.33 (± 2.21)	1.49 (± 2.85)	97.73 (± 1.32)	1.39 (± 0.82)
Proposed	**1.90 (± 1.77)**	**0.21 (± 2.58)**	**97.76 (± 1.32)**	**1.32 (± 0.73)**

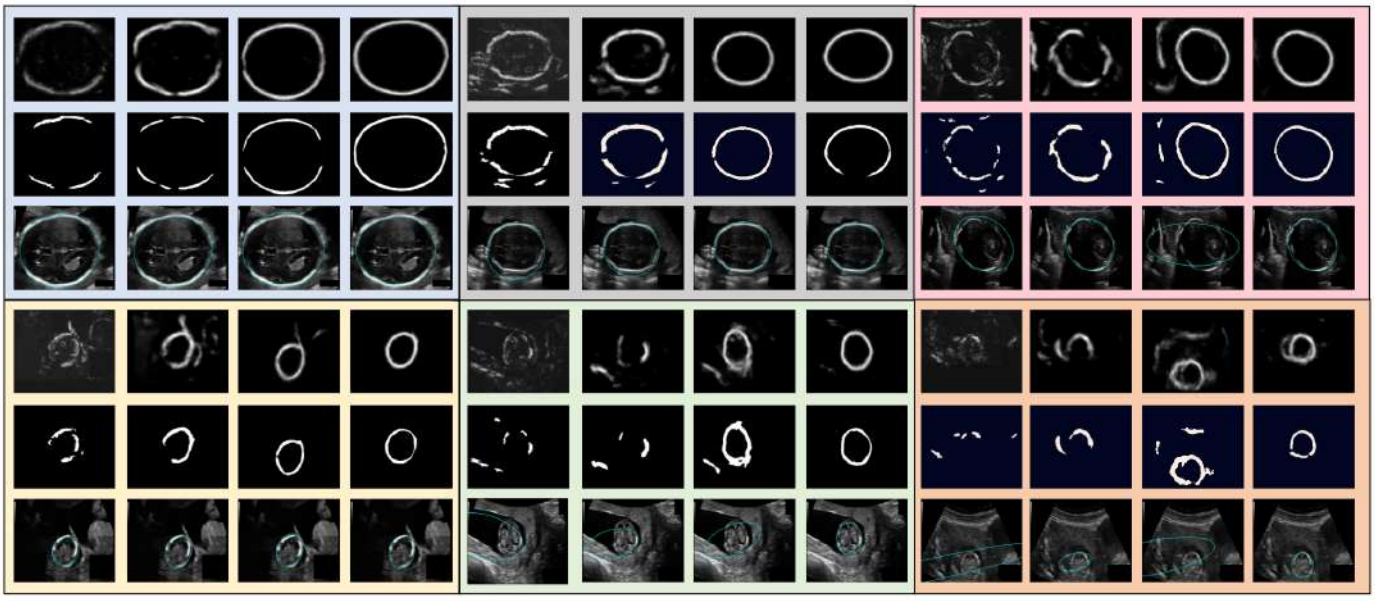


Fig. 9: Visual samples of the ablation-study results. Each colored panel corresponds to a test image. Each panel shows the results obtained by (1st column) shallow regression, (2nd column) short proposed UNet, (3rd column) regression only and (4th column) proposed framework. Each column shows the (first row) prediction, (second row) post processed and (third-row) ellipse-fitted image.

a momentum of 0.98, for 1500 epochs. The value of the momentum was set according to similar work that uses regression networks [5]. Batch size was set to 32, as a trade off between memory requirements and training convergence. On-the-fly data augmentation was performed during training. Augmentation included horizontal flip and rotation in range ±45°, and was randomly applied at each training iteration. The best model among epochs was selected according to the lowest mean absolute error obtained in the validation set. The threshold value was set to the 80% of the maximum value of the regression output, and the outliers for the post-processing described in Sec. 3.3 were identified using a O value of 0.5. The empirical threshold value of 80% and the O value of 0.5 were optimized on the validation set.

The HC length [mm] was obtained by multiplying the HC pixels for the corresponding pixel size [mm], provided by the challenge organizers. To obtain the final HC parameters, all the images were resized and recentered.

All the analyses were performed using the *Keras*[4] Python library on a NVIDIA RTX 2080TI, with a Xeon e5 CPU and 128 GB RAM.

4.3. Ablation study and comparison with the state of the art

As a first ablation study, we investigated the use of a simple regression architecture (shallow regression). The shallow regression only architecture was chosen following similar work in the literature [5].

The network was composed of 6 convolutional layers followed by batch normalization and activated with ReLU. We also tested UNet with only the 2 top layers at first (proposed-short) and then with an additional top layer (proposed-middle) (plus the bottleneck in both the architectures) to investigate the trade off between architecture dept hand HC delineation performance. Both shallow regression, proposed-short and proposed-middle results are obtained including the post-processing steps presented in Sec. 3.3

To assess whether the fetal-head localization and centering helped the regression network, we compared the performance achieved by the proposed framework with that achieved by the regression network fed with raw US images.

We also tested the performance of our network without excluding outliers.

For fair comparison, all approaches were investigated using the same dataset split and training setting, as well as the same computational hardware.

For the comparison with the state of the art, we chose the methods proposed in [25], [28], [30] and [34], which follow the deep-learning paradigm and are the most similar with respect to our approach. All these methods were developed and tested using the *HC18 Grand Challenge* dataset. The relative performance metrics, reported in Sec. 5, were extracted by the corresponding research papers. We also included the work in [19], even if it uses handcrafted features, because it is the one that presented the *HC18 Grand Challenge* dataset.

[4] https://keras.io/

4.4. Performance metrics

To measure the performance of the proposed method, we computed the metrics suggested by the organizers of the *HC18 Challenge*. Hence, the difference (DF) [mm], absolute difference (ADF) [mm], Hausdorff distance (HD) [mm] and Dice similarity coefficient (DSC) [%] were computed.

The DF and ADF are defined as follows:

$$DF = HC_A - HC_B \tag{2}$$

$$ADF = |HC_A - HC_B| \tag{3}$$

with HC_A and HC_B the HC obtained by the proposed method and clinician manual annotation, respectively.

The HD is computed as

$$HD(A, B) = max(h(A, B), h(A, B)) \tag{4}$$

where $B = \{b_1, ..., b_q\}$ and $A = \{a_1, ..., a_q\}$ are the sets of pixels from the HC measured by clinician and obtained with our framework, respectively, with

$$h(A, B) = \max_{a \in A} \max_{b \in B} \|a - b\| \tag{5}$$

The DSC is defined as:

$$DSC = \frac{2 * |Area_a \cap Area_b|}{|Area_a| + |Area_b|} \tag{6}$$

with $Area_b$ and $Area_a$ the fetal head area as delimited by the HC measured by clinicians and obtained with the proposed framework, respectively.

5. Results

The performance comparison in terms of ADF, DF, DSC and HD of the different models proposed in the ablation study is summarized in Table 2. The lowest mean AD = 1.90 (± 1.77) mm was obtained with the proposed framework, with a mean AD of 1.49 (± 1.32), 1.72 (± 1.50) and 3.25 (± 2.64) mm for images in the first, second and third trimester, respectively. The worst results were achieved with the shallow regression network, with an AD of 11.11 (± 38.25) mm. With the proposed-short and the regression-only architecture, the AD dropped to 5.67 (± 30.28) and 5.79 (± 16.58) mm, respectively. Proposed-middle achieved good results, with a mean AD of 2.08 (± 2.08). The same trend was observed for the other metrics. Without excluding the outliers from the distance distribution, as described in Sec. 3.3, the results of the proposed framework slightly dropped to (AD) 2.33 (± 3.36) mm, (DF) 1.46 (± 4.10) mm, (DSC) 97.30 (± 3.21)%, (HD) 1.51 (± 3.68) mm.

Figure 9 shows visual samples of the results obtained from different models in the ablation study. Each colored panel refers to a test image. The first, second, third and fourth columns shows the results achieved by shallow regression, proposed-short, regression only and proposed framework, respectively. The upper, middle and lower images correspond to prediction, post-processed and ellipse-fitted images, respectively. Since no

radical differences were found by visual inspection, it was decided not to include proposed-middle and Proposed without excluding outliers in the figure.

As shown in the blue panel, each of the models performed comparably when processing images with clearly-contrasted, centered HC that covers a large portion of the image. As the dimension of the HC decreased (gray panel), the shallow regression and easy UNet predictions deteriorated. The pink panel shows an image where noise and blur make the HC poorly visible with respect to the background, resulting in low performance for the shallow regression, easy UNet and regression-only architecture. The yellow and green box shows challenging images (poorly visible head skull and evident uterus edge) that compromised even further the prediction of the shallow regression, easy UNet and regression-only network. The post processing helped to attenuate the prediction inaccuracies, as shown in the yellow panel, while it was not effective when the HC texture was similar to the background tissues.

In the test image shown in the orange panel, only the proposed framework was able to delineate the head quite accurately.

Moving to the comparison with the state of the art, the ADF, DF, DSC and HD obtained by work in literature that used the $HC18$ *Grand Challenge* are reported in Table 3. The best performance, for all metrics, was achieved by the proposed framework. The work in [28] and [30] achieved ADs very close to ours, with a difference of 0.22 and 0.32 mm, respectively.

Den Heuvel et al.[19], Rong et al.[25] and Al-Bander et al.[34] obtained the largest ADs.

The inference time of the entire pipeline was of ~0.03s when using the NVIDIA RTX 2080TI with a Xeon e5 CPU and 128 GB RAM. For the sake of completeness, we also computed the inference time on a less performing machine (a MacBook Pro with a Intel Iris Plus Graphics 1536 MB with 16 GB RAM), achieving an inference of 0.7s.

6. Discussion

Measuring fetal biometrics, such as HC length, has a strong diagnostic and prognostic role. In the actual clinical practice, biometric measurements are performed manually, resulting in high intra- and inter-clinician variability. Computer-assisted solutions in the literature try to tackle this issue by proposing approaches based on deep learning for fetal head segmentation. In this work, we instead investigated if it was possible to directly regress a distance field from the HC for accurate HC delineation. To relieve the regression network of learning the HC position in the image, we exploited a region-proposal network to detect and center HC before feeding the regression network. Directly regressing HC boundary is empirically too localized (i.e., it supports small spatial context) therefore we avoided to do it merely with the pixel positions. We instead regressed distance fields, as shown to be successful in the literature of closer branch studies (e.g., [6, 8]). We tested the proposed framework on the *HC18 Grand Challenge*, in which there are images with different HC dimension and location, poorly visible HC as opposed to the background, and presence of reverberations, shadows and speckles.

The proposed framework achieved encouraging results with a mean AD of 1.90 mm, showing also high stability, as demonstrated by low standard-deviation values (± 1.76). The value obtained is two order of magnitude lower with respect to the HC length (mean HC in the training set = 174.38 mm), thus showing a great potential for clinical practice applications. A slightly lower performance was seen for images from the third trimester (with a difference of 1.76 mm between the AD of the first and third trimesters), which may be due to the higher pixel dimension in images from the third semester with respect to images from the other trimesters.

The combination between the detection and regression networks allowed to obtain accurate predictions, which were not affected by the fetal head position in the image. This was proven by the lowest performance of the regression-only network, where the network had also to learn to retrieve the position of the HC in the images (see Fig. 9, pink, green and orange panels). This is particularly evident from the comparison with the Proposed without excluding outlier: the proposed method turns out to be better (AD of 2.3 mm) even without discarding outliers. It is worth highlighting though how ours is just one of the many approaches that can be used to exclude outliers. Other viable options include connected component analysis.

As the depth increased, the regression network was able to discriminate characterizing head features with respect to features of other structures (e.g., uterus edges) with similar intensity level (see Fig. 9, green and orange panels). This was not seen in fact when testing the shallow regression and the Proposed-short networks, as opposed to Proposed-middle and Proposed.

The proposed framework provided the highest performance among the state of the art methods tested on the same dataset. The lower performance of [19] with respect to our approach can be explained considering the highest robustness of deep learning over handcrafted-based approaches. Similarly, our data-drive approach showed to be more robust with respect to the model-based strategy of [25]. A possible reason for the lower performance of [28] may be attributed to the challenging task of directly regressing the HC parameters. Our approach also outperformed approaches based on fetal-head segmentation [34], [30]), showing that modeling the HC delineation as a edge-delineation problem, by directly regressing a distance field from the HC, may be a valuable alternative for HC length computation.

The proposed pipeline achieved real-time inference using a powerful GPU. At the same time, when using a less powerful computational resource, we achieved an inference time lower than 1 s. We believe this may be suitable for clinical applications, considering that the manual selection of point on the US image by clinicians could take more than 1 s.

A straightforward limitation of this work may be seen in the size of the dataset, which, however, is considered as benchmark in the field. Furthermore, all the data used in the study were acquired from only two different devices (same vendor) in the same hospital, and only one clinician performed the HC annotation [19]. With a view to translate the proposed pipeline in the actual clinical practice, more data are required for validation

proposes.

Future directions of this work include exploring feedback from users, as recently proposed in [33]. Moreover, visual attention mechanism could be encoded in the regression network for further boosting the results [41].

We also would like to investigate more advanced models such as nnUNet [42], for improving the delineation performance. Adversarial training to take into account HC shape priors may also be explored [43], as well as atlas-based approaches [44]. As future work, we also plan to exploit synthetic augmentation techniques through generative adversarial networks. Image synthesis has been shown to be a valuable tool for improving performances, and may help to reduce the errors due to the poor generalisation of rare patterns in the training set [45].

7. Conclusions

In this work, we showed that exploiting regression CNNs for HC delineation from US images may be a valuable solution to automatically generate HC measurements. We achieved an $AD = 1.90\ (\pm\ 1.76)$ mm, overcoming the approaches presented in the literature. Inspired by the excellent performance of boundary detection-based algorithms [46] [47], this work is among the first attempts of exploiting regression CNNs for edge-localization tasks in the US field and has great potential to support clinicians during the clinical practice for fetal biometric measurement.

Acknowledgments

This study did not need any ethical approval. The authors declare no competing interests.

References

[1] S. Degani, Fetal biometry: clinical, pathological, and technical considerations, Obstetrical & Gynecological Survey 56 (3) (2001) 159–167.

[2] B. Deloison, G. E. Chalouhi, J.-P. Bernard, Y. Ville, L. J. Salomon, Outcomes of fetuses with small head circumference on second-trimester ultrasonography, Prenatal Diagnosis 32 (9) (2012) 869–874.

[3] J. Espinoza, S. Good, E. Russell, W. Lee, Does the use of automated fetal biometry improve clinical work flow efficiency?, Journal of Ultrasound in Medicine 32 (5) (2013) 847–850.

[4] S. Rueda, S. Fathima, C. L. Knight, M. Yaqub, A. T. Papageorghiou, B. Rahmatullah, A. Foi, M. Maggioni, A. Pepe, J. Tohka, et al., Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge, IEEE Transactions on Medical Imaging 33 (4) (2013) 797–813.

[5] X. Du, T. Kurmann, P.-L. Chang, M. Allan, S. Ourselin, R. Sznitman, J. D. Kelly, D. Stoyanov, Articulated multi-instrument 2-D pose estimation using fully convolutional networks, IEEE Transactions on Medical Imaging 37 (5) (2018) 1276–1287.

[6] E. Colleoni, S. Moccia, X. Du, E. De Momi, D. Stoyanov, Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers, IEEE Robotics and Automation Letters 4 (3) (2019) 2714–2721.

[7] G. Hattab, M. Arnold, L. Strenger, M. Allan, D. Arsentjeva, O. Gold, T. Simpfendörfer, L. Maier-Hein, S. Speidel, Kidney edge detection in laparoscopic image data for computer-assisted surgery, International Journal of Computer Assisted Radiology and Surgery (2019) 1–9.

[8] S. Moccia, L. Migliorelli, V. Carnielli, E. Frontoni, Preterm infants' pose estimation with spatio-temporal features, IEEE Transactions on Biomedical Engineering (2019).

[9] I. P. Satwika, I. Habibie, M. A. Ma'sum, A. Febrian, E. Budianto, Particle swarm optimation based 2-dimensional randomized Hough transform for fetal head biometry detection and approximation in ultrasound imaging, in: International Conference on Advanced Computer Science and Information System, IEEE, 2014, pp. 468–473.

[10] W. Lu, J. Tan, Detection of incomplete ellipse in images with strong noise by iterative randomized Hough transform (IRHT), Pattern Recognition 41 (4) (2008) 1268–1279.

[11] C. Sun, Automatic fetal head measurements from ultrasound images using circular shortest paths, IEEE International Symposium on Biomedical Imaging, Challenge US: Biometric Measurements from Fetal Ultrasound Images (2012) 13–15.

[12] A. Foi, M. Maggioni, A. Pepe, J. Tohka, Head contour extraction from the fetal ultrasound images by difference of Gaussians revolved along elliptical paths, IEEE International Symposium on Biomedical Imaging, Challenge US: Biometric Measurements from Fetal Ultrasound Images (2012) 1–3.

[13] J. Perez-Gonzalez, J. B. Muñoz, M. R. Porras, F. Arámbula-Cosío, V. Medina-Bañuelos, Automatic fetal head measurements from ultrasound images using optimal ellipse detection and texture maps, in: VI Latin American Congress on Biomedical Engineering, Springer, 2015, pp. 329–332.

[14] V. Rajinikanth, N. Dey, R. Kumar, J. Panneerselvam, N. S. M. Raja, Fetal head periphery extraction from ultrasound image using Jaya algorithm and Chan-Vese segmentation, Procedia Computer Science 152 (2019) 66–73.

[15] R. V. Stebbing, J. E. McManigle, A boundary fragment model for head segmentation in fetal ultrasound, Challenge US: Biometric Measurements from Fetal Ultrasound Images (2012) 9–11.

[16] E. A. Anto, B. Amoah, A. Crimi, Segmentation of ultrasound images of fetal anatomic structures using random forest for low-cost settings, in: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2015, pp. 793–796.

[17] G. Carneiro, B. Georgescu, S. Good, D. Comaniciu, Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree, IEEE Transactions on Medical Imaging 27 (9) (2008) 1342–1355.

[18] J. Li, Y. Wang, B. Lei, J.-Z. Cheng, J. Qin, T. Wang, S. Li, D. Ni, Automatic fetal head circumference measurement in ultrasound using Random forest and fast ellipse fitting, IEEE Journal of Biomedical and Health Informatics 22 (1) (2017) 215–223.

[19] T. L. van den Heuvel, D. de Bruijn, C. L. de Korte, B. van Ginneken, Automated measurement of fetal head circumference using 2D ultrasound images, PloS One 13 (8) (2018) e0200412.

[20] L. Zhang, X. Ye, T. Lambrou, W. Duan, N. Allinson, N. J. Dudley, A supervised texton based approach for automatic segmentation and measurement of the fetal head and femur in 2D ultrasound images, Physics in Medicine & Biology 61 (3) (2016) 1095.

[21] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88.

[22] P. Zaffino, G. Pernelle, A. Mastmeyer, A. Mehrtash, H. Zhang, R. Kikinis, T. Kapur, M. F. Spadea, Fully automatic catheter segmentation in MRI with 3D convolutional neural networks: application to MRI-guided gynecologic brachytherapy, Physics in Medicine & Biology 64 (16) (2019) 165008.

[23] R. Rosati, L. Romeo, S. Silvestri, F. Marcheggiani, L. Tiano, E. Frontoni, Faster r-cnn approach for detection and quantification of dna damage in comet assay images, Computers in Biology and Medicine (2020) 103912.

[24] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, T. Wang, Deep learning in medical ultrasound analysis: a review, Engineering (2019).

[25] Y. Rong, D. Xiang, W. Zhu, F. Shi, E. Gao, Z. Fan, X. Chen, Deriving external forces via convolutional neural networks for biomedical image segmentation, Biomedical Optics Express 10 (8) (2019) 3800–3814.

[26] A. Fitzgibbon, M. Pilu, R. B. Fisher, Direct least square fitting of ellipses, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (5) (1999) 476–480.

[27] K. Irene, H. Haidi, N. Faza, W. Chandra, et al., Fetal head and abdomen measurement using convolutional neural network, Hough transform, and difference of Gaussian revolved along elliptical path (Dogell) algorithm, Available online at: arXiv preprint arXiv:1911.06298 (2019).

[28] Z. Sobhaninia, S. Rafiei, A. Emami, N. Karimi, K. Najarian, S. Samavi, S. R. Soroushmehr, Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning, in: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2019, pp. 6545–6548.
[29] A. Chaurasia, E. Culurciello, Linknet: Exploiting encoder representations for efficient semantic segmentation, in: IEEE Visual Communications and Image Processing, IEEE, 2017, pp. 1–4.
[30] Z. Sobhaninia, A. Emami, N. Karimi, S. Samavi, Localization of fetal head in ultrasound images by multiscale view and deep neural networks, Available online at: arXiv preprint arXiv:1911.00908 (2019).
[31] H. P. Kim, S. M. Lee, J.-Y. Kwon, Y. Park, K. C. Kim, J. K. Seo, Automatic evaluation of fetal head biometry from ultrasound images using machine learning, Physiological Measurement (2019).
[32] J. J. Cerrolaza, M. Sinclair, Y. Li, A. Gomez, E. Ferrante, J. Matthew, C. Gupta, C. L. Knight, D. Rueckert, Deep learning with ultrasound physics for fetal skull segmentation, in: IEEE International Symposium on Biomedical Imaging, IEEE, 2018, pp. 564–567.
[33] S. Budd, M. Sinclair, B. Khanal, J. Matthew, D. Lloyd, A. Gomez, N. Toussaint, E. C. Robinson, B. Kainz, Confident head circumference measurement from ultrasound with real-time feedback for sonographers, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 683–691.
[34] B. Al-Bander, T. Alzahrani, S. Alzahrani, B. M. Williams, Y. Zheng, Improving fetal head contour detection by object localisation with deep learning, in: Annual Conference on Medical Image Understanding and Analysis, Springer, 2019, pp. 142–150.
[35] A. Bulat, G. Tzimiropoulos, Human pose estimation via convolutional part heatmap regression, in: European Conference on Computer Vision, Springer, 2016, pp. 717–732.
[36] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.
[37] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[38] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, X. Ding, Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation, Medical Image Analysis (2020) 101693.
[39] P. J. Rousseeuw, M. Hubert, Anomaly detection by robust statistics, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 (2) (2018) e1236.
[40] D. K. Prasad, M. K. Leung, C. Quek, Ellifit: An unconstrained, non-iterative, least squares based geometric ellipse fitting method, Pattern Recognition 46 (5) (2013) 1449–1465.
[41] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, Attention gated networks: Learning to leverage salient regions in medical images, Medical Image Analysis 53 (2019) 197–207.
[42] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, et al., nnu-net: Self-adapting framework for u-net-based medical image segmentation, arXiv preprint arXiv:1809.10486 (2018).
[43] A. Casella, S. Moccia, E. Frontoni, D. Paladini, E. De Momi, L. S. Mattos, Inter-foetus membrane segmentation for TTTS using adversarial networks, Annals of Biomedical Engineering 48 (2) (2020) 848–859.
[44] P. Zaffino, P. Raudaschl, K. Fritscher, G. C. Sharp, M. F. Spadea, Plastimatch MABS, an open source tool for automatic image segmentation, Medical p Physics 43 (9) (2016) 5155–5160.
[45] F. Calimeri, A. Marzullo, C. Stamile, G. Terracina, Biomedical data augmentation using generative adversarial neural networks, in: International Conference on Artificial Neural Networks, Springer, 2017, pp. 626–634.
[46] S. Yin, Q. Peng, H. Li, Z. Zhang, X. You, K. Fischer, S. L. Furth, G. E. Tasian, Y. Fan, Automatic kidney segmentation in ultrasound images using subsequent boundary distance regression and pixelwise classification networks, Medical image analysis 60 (2020) 101602.
[47] H. Tang, M. Moradi, K. C. Wong, H. Wang, A. El Harouni, T. Syeda-Mahmood, Integrating deformable modeling with 3d deep neural network segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 377–384.

Vitae

Maria Chiara Fiorentino (B.Sc. 2015, M.Sc, 2018) was born in Ancona (AN) on June 1992. She graduated cum laude in Biomedical Engineering at Università Politecnica delle Marche (Ancona, Italy) in July 2018, with a thesis entitled: "Design and implementation of a dyskinesia detection system for patients with Parkinson's disease based on consumer electronics devices". Between 2018 and 2019, she spent 6 months at the Leiden University Medical Center (LUMC) as research fellows. The research was focused on the development of advanced image-processing methods for image segmentation in cardiac images. Maria Chiara is currently a PhD at Università Politecnica delle Marche, in the Department of Information Engineering. Her PhD project deals with deep-learning methods for ultrasound image analysis.

Sara Moccia (B.Sc. 2012, M.Sc. 2014, Ph.D. 2018) was born in Bari (BA) on September 1990. She graduated cum laude in Biomedical Engineering at Politecnico di Milano (Milan, Italy) in December 2014, with a thesis entitled: "Statistical-segmentation techniques of liver metastases and necroses in FGD-PET for the automatic evaluation of pre and post thermoablation PET/CT studies". In May 2018, she obtained the European PhD cum laude in Bioengineering from Istituto Italiano di Tecnologia (Genoa, Italy) and Politecnico di Milano with a thesis entitled "Supervised tissue classification in optical images: Towards new applications of surgical data science". During her PhD, she was hosted at the Computer-Assisted Medical Interventions laboratory at the German Cancer Research Center (Heidelberg, Germany). Sara is currently Postdoc at Università Politecnica delle Marche (Ancona, Italy) and Research Fellow at Istituto Italiano di Tencologia.

Morris Capparuccini (B.Sc. 2018) was born in Fermo (FM) on December 1990. He graduated in Computer Engineering and Automation at Università Politecnica delle Marche (Ancona, Italy) in July 2018, with a thesis entitled: "Development of a gaming application based on gesture recognition". Morris is currently a M.Sc. student in Computer Engineering and Automation at Università Politecnica delle Marche.

Sara Giamberini (B.Sc. 2018) was born in Ancona (AN) on March 1993. She graduated in Computer Engineering and Automation at Università Politecnica delle Marche (Ancona, Italy) in October 2018, with a thesis entitled: "A multi-objective mathematical programming based approach for querying Data Warehouse". Sara is currently a M.Sc. student in Computer Engineering and Automation at Università Politecnica delle Marche.

Emanuele Frontoni (M.Sc 2002, Ph.D 2006) was born in Fermo (FM) on May 1978. He graduated cum laude in Electronic Engineering at Università Politecnica delle Marche (Ancona, Italy) in 2002, with a thesis entitled: "Methods for extracting information from DNA sequences". He obtained the PhD in Intelligent Artificial Systems from the Department of In-

formation, Business and Automation Engineering of Università Politecnica delle Marche, discussing a thesis on Vision Based Robotics. His research focuses on applying computer science, artificial intelligence and computer vision techniques to mobile robots, innovative IT applications and medical data. He is a member of IEEE and AI*IA, the Italian Association for Artificial Intelligence.