

Real-Time Standard Scan Plane Detection and Localisation in Fetal Ultrasound Using Fully Convolutional Neural Networks

Christian F. Baumgartner¹(✉), Konstantinos Kamnitsas¹,
Jacqueline Matthew^{2,3}, Sandra Smith³, Bernhard Kainz¹, and Daniel Rueckert¹

¹ Biomedical Image Analysis Group, Imperial College London, London, UK
c.baumgartner@imperial.ac.uk

² Biomedical Research Centre, Guy's and St Thomas' NHS Foundation, London, UK

³ Division of Imaging Sciences and Biomedical Engineering,
King's College London, London, UK

Abstract. Fetal mid-pregnancy scans are typically carried out according to fixed protocols. Accurate detection of abnormalities and correct biometric measurements hinge on the correct acquisition of clearly defined standard scan planes. Locating these standard planes requires a high level of expertise. However, there is a worldwide shortage of expert sonographers. In this paper, we consider a fully automated system based on convolutional neural networks which can detect twelve standard scan planes as defined by the UK fetal abnormality screening programme. The network design allows real-time inference and can be naturally extended to provide an approximate localisation of the fetal anatomy in the image. Such a framework can be used to automate or assist with scan plane selection, or for the retrospective retrieval of scan planes from recorded videos. The method is evaluated on a large database of 1003 volunteer mid-pregnancy scans. We show that standard planes acquired in a clinical scenario are robustly detected with a precision and recall of 69% and 80%, which is superior to the current state-of-the-art. Furthermore, we show that it can retrospectively retrieve correct scan planes with an accuracy of 71% for cardiac views and 81% for non-cardiac views.

1 Introduction

Abnormal fetal development is a leading cause of perinatal mortality in both industrialised and developing countries [11]. Although many countries have introduced fetal screening programmes based on mid-pregnancy ultrasound (US) scans at around 20 weeks of gestational age, detection rates remain relatively low. For example, it is estimated that in the UK approximately 26% of fetal anomalies are not detected during pregnancy [4]. Detection rates have also been reported to vary considerably across different institutions [1] which suggests that, at least in part, differences in training may be responsible for this variability. Moreover, according to the WHO, it is likely that worldwide many US scans are carried out by individuals with little or no formal training [11].

Biometric measurements and identification of abnormalities are performed on a number of standardised 2D US view planes acquired at different locations in the fetal body. In the UK, guidelines for selecting these planes are defined in [7]. Standard scan planes are often hard to localise even for experienced sonographers and have been shown to suffer from low reproducibility and large operator bias [4]. Thus, a system automating or aiding with this step could have significant clinical impact particularly in geographic regions where few highly skilled sonographers are available. It is also an essential step for further processing such as automated measurements or automated detection of anomalies.

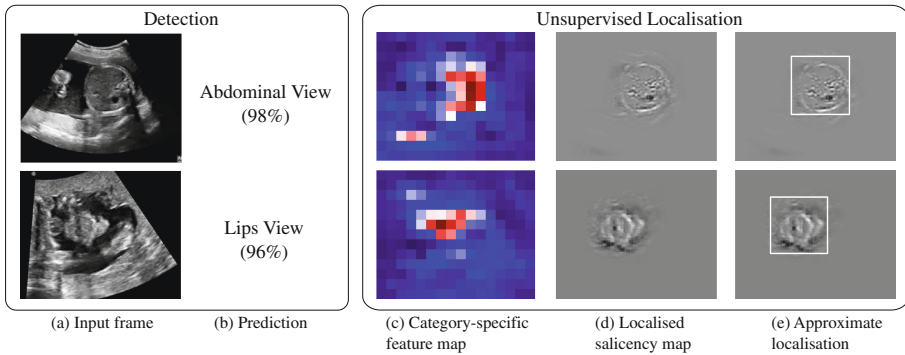


Fig. 1. Overview of the proposed framework for two standard view examples. Given a video frame (a) the trained convolutional neural network provides a prediction and confidence value (b). By design, each classifier output has a corresponding low-resolution feature map (c). Back-propagating the error from the most active feature neurons results in a saliency map (d). A bounding box can be derived using thresholding (e).

Contributions: We propose a real-time system which can automatically detect 12 commonly acquired standard scan planes in clinical free-hand 2D US data. We demonstrate the detection framework for (1) real-time annotations of US data to assist sonographers, and (2) for the retrospective retrieval of standard scan planes from recordings of the full examination. The method employs a fully convolutional neural network (CNN) architecture which allows robust scan plane detection at more than 100 frames per second. Furthermore, we extend this architecture to obtain saliency maps highlighting the part of the image that provides the highest contribution to a prediction (see Fig. 1). Such saliency maps provide a localisation of the respective fetal anatomy and can be used as starting point for further automatic processing. This localisation step is unsupervised and does not require ground-truth bounding box annotations during training.

Related Work: Standard scan plane classification of 7 planes was proposed for a large fetal *image* database [13]. This differs significantly from our work since in that scenario it is already known that every image is in fact a standard plane whilst in video data the majority of frames does not show standard planes.

A number of papers have proposed methods to detect fetal anatomy in videos of fetal 2D US sweeps (e.g. [6]). In those works the authors were aiming at detecting the presence of fetal structures such as the skull, heart or abdomen rather than specific standardised scan planes. Automated fetal standard scan plane detection has been demonstrated for 1–3 standard planes in 2D fetal US sweeps [2, 3, 8]. Notably, [2, 3] also employed CNNs. US sweeps are acquired by moving the US probe from the cervix upwards in one continuous motion [3]. However, not all standard views required to determine the fetus’ health status are adequately visualised using a sweep protocol. For example, visualising the femur or the lips normally requires careful manual scan plane selection. Furthermore, data obtained using the sweep protocol are typically only 2–5 s long and consist of fewer than 50 frames [3]. To the best of our knowledge, fetal standard scan plane detection has never been performed on true free-hand US data which typically consist of 10,000+ frames. Moreover, none of related works were demonstrated to run in real-time, typically requiring multiple seconds per frame.

2 Materials and Methods

Data and Preprocessing: Our dataset consists of 1003 2D US scans of consented volunteers with gestational ages between 18–22 weeks which have been acquired by a team of expert sonographers using GE Voluson E8 systems. For each scan a screen capture video of the entire procedure was recorded. Additionally, the sonographers saved “freeze frames” of a number of standard views for each subject. A large fraction of these frames have been annotated allowing us to infer the correct ground-truth (GT) label. All video frames and images were downsampled to a size of 225×273 pixels.

We considered 12 standard scan planes based on the guidelines in [7]. In particular, we selected the following: two brain views at the level of the ventricles (Vt.) and the cerebellum (Cb.), the standard abdominal view, the transverse kidney view, the coronal lip, the median profile, and the femur and sagittal spine views. We also included four commonly acquired cardiac views: the left and right ventricular outflow tracts (LVOT and RVOT), the three vessel view (3VV) and the 4 chamber view (4CH)¹. In addition to the labelled freeze frames, we sampled 50 random frames from each video in order to model the background class, i.e., the “not a standard scan plane” class.

Network Architecture: The architecture of our proposed CNN is summarised in Fig. 2. Following recent advances in computer vision, we opted for a fully convolutional network architecture which replaces traditional fully connected layers with convolution layers using a 1×1 kernel [5, 9]. In the final convolutional layer (C6) the input is reduced to K 13×13 feature maps F_k , where K is the number of classes. Each of these feature maps is then averaged to obtain the

¹ A detailed description of the considered standard planes is included in the supplementary material available at <http://www.doc.ic.ac.uk/~cbaumgar/dwnlds/miccai2016/>.

input to the final Softmax layer. This architecture makes the network flexible with regard to the size of the input images. Larger images will simply result in larger feature maps, which will nevertheless be mapped to a scalar for the final network output. We use this fact to train on cropped square images rather than the full field of view which is beneficial for data augmentation.

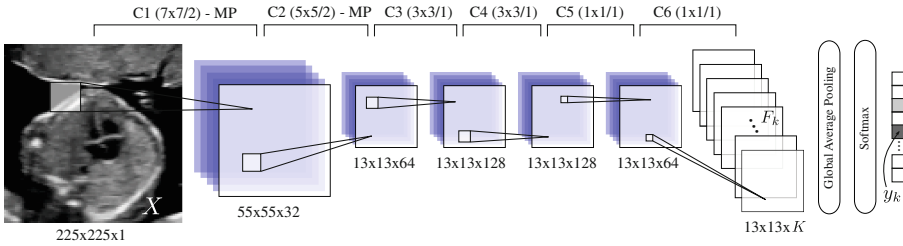


Fig. 2. Overview of the proposed network architecture. The size and stride of the convolutional kernels are indicated at the top (notation: $kernel\ size/stride$). Max-pooling steps are indicated by *MP* (2×2 bins, stride of 2). The activation functions of all convolutions except *C6* are rectified non-linear units (ReLU). *C6* is followed by a global average pooling step. The sizes at the bottom of each image/feature map refer to the training phase and will be slightly larger during inference due to larger input images.

A key aspect of our proposed network architecture is that we enforce a one-to-one correspondence between each feature map F_k and the respective prediction y_k . Since each neuron in the feature maps F_k has a receptive field in the original image, during training, the neurons will learn to activate only if an object of class k is in that field. This allows to interpret F_k as a spatially encoded confidence map for class k [5]. In this paper, we take advantage of this fact to generate localised saliency maps as described below.

Training: We split our dataset into a test set containing 20% of the subjects and a training set containing 80%. We use 10% of the training data as validation set to monitor the training progress. In total, we model 12 standard view planes, plus one background class resulting in $K = 13$ categories.

We train the model using mini-batch gradient descent and the categorical cross-entropy cost function. In order to prevent overfitting we add 50% dropout after the C5 and C6 layers. To account for the significant class imbalance introduced by the background category, we create mini-batches with even class-sampling. Additionally, we augment each batch by a factor of 5 by taking 225×225 square sub-images with a random horizontal translation and transforming them with a small random rotation and flips along the vertical axis. Taking random square sub-images allows to introduce more variation to the augmented batches compared to training on the full field of view. This helps to reduce the overfitting of the network. We train the network for 50 epochs and choose the network parameters with the lowest error on the validation set.

Frame Annotation and Retrospective Retrieval: After training we feed the network with video frames containing the full field of view (225×273 pixels) of the input videos. This results in larger category-specific feature maps of 13×16 . The prediction y_k and confidence c_k of each frame are given by the prediction with the highest probability and the probability itself.

For retrospective frame retrieval, for each subject we calculate and record the confidence for each class over the entire duration of an input video. Subsequently, we retrieve the frame with the highest confidence for each class.

Saliency Maps and Unsupervised Localisation: After obtaining the category y_k of the current frame X from a forward pass through the network, we can examine the feature map F_k (i.e. the output of the C6 layer) corresponding to the predicted category k . Two examples of feature maps are shown in Fig. 1c. The F_k could already be used to make an approximate estimate of the location of the respective anatomy similar to [9].

Here, instead of using the feature maps directly, we present a novel method to obtain localised saliency with the resolution of the original input images. For each neuron $F_k^{(p,q)}$ at the location p, q in the feature map it is possible calculate how much each original input pixel $X^{(i,j)}$ contributed to the activation of this neuron. This corresponds to calculating the partial derivatives

$$S_k^{(i,j)} = \frac{\partial F_k^{(p,q)}}{\partial X^{(i,j)}},$$

which can be solved efficiently using an additional backwards pass through the network. [12] proposed a method for performing this back-propagation in a *guided* manner by allowing only error signals which contribute to an increase of the activations in the higher layers (i.e. layers closer to the network output) to back-propagate. In particular, the error is only back-propagated through each neuron's ReLU unit if the input to the neuron x , as well as the error in the higher layer δ_ℓ are positive. That is, the back-propagated error $\delta_{\ell-1}$ of each neuron is given by $\delta_{\ell-1} = \delta_\ell \sigma(x) \sigma(\delta_\ell)$, where $\sigma(\cdot)$ is the unit step function.

In contrast to [12] who back-propagated from the final output, in this work we take advantage of the spatial encoding in the category specific feature maps and only back-propagate the errors for the 10% most active feature map neurons, i.e. the spatial locations where the fetal anatomy is predicted. The resulting saliency maps are significantly more localised compared to [12] (see Fig. 3).

These saliency maps can be used as starting point for various image analysis tasks such as automated segmentation or measurements. Here, we demonstrate how they can be used for approximate localisation using basic image processing. We blur the absolute value image of a saliency map $|S_k|$ using a 25×25 Gaussian kernel and apply a thresholding using Otsu's method [10]. Finally, we compute the minimum bounding box of the components in the thresholded image.

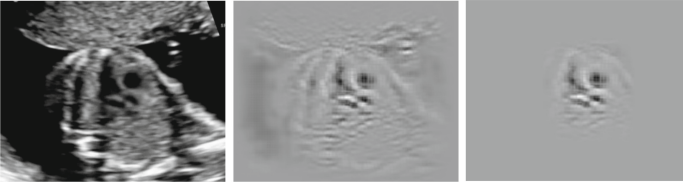


Fig. 3. Saliency maps obtained from the input frame (LVOT class) shown on the *left*. The *middle* map was obtained using guided back-propagation from the average pool layer output [12]. The map on the *right* was obtained using our proposed method.

3 Experiments and Results

Frame Annotation: We evaluated the ability of our method to detect standard frames by classifying the test data including the randomly sampled background class. We report the achieved precision (pc) and recall (rc) scores in Table 1. The lowest scores were obtained for cardiac views, which are also the most difficult to scan for expert sonographers. This fact is reflected in the low detection rates for serious cardiac anomalies (e.g. only 35 % in the UK).

Table 1. Precision $pc = TP/(TP + FP)$ and recall $rc = TP/(TP + FN)$ for the classification of the modelled scan planes. **Background class:** $pc = 0.96$, $rc = 0.93$.

view	pc	rc	view	pc	rc	view	pc	rc
Brain (Vt.)	0.96	0.90	Lips	0.85	0.88	LVOT	0.63	0.63
Brain (Cb.)	0.92	0.94	Profile	0.71	0.82	RVOT	0.40	0.46
Abdominal	0.85	0.80	Femur	0.79	0.93	3VV	0.46	0.60
Kidneys	0.64	0.87	Spine	0.51	0.99	4CH	0.61	0.74

[2] have recently reported pc/rc scores of 0.75/0.75 for the abdominal standard view, and 0.77/0.61 for the 4CH view in US sweep data. We obtained comparable values for the 4CH view and considerably better values for the abdominal view. However, with 12 modelled standard planes and free-hand US data

Table 2. % of correctly retrieved frames for each standard view for all 201 test subjects.

view	%	view	%	view	%
Brain (Vt.)	0.95	Lips	0.77	LVOT	0.73
Brain (Cb.)	0.89	Profile	0.76	RVOT	0.70
Abdominal	0.79	Femur	0.75	3VV	0.66
Kidneys	0.87	Spine	0.77	4CH	0.78

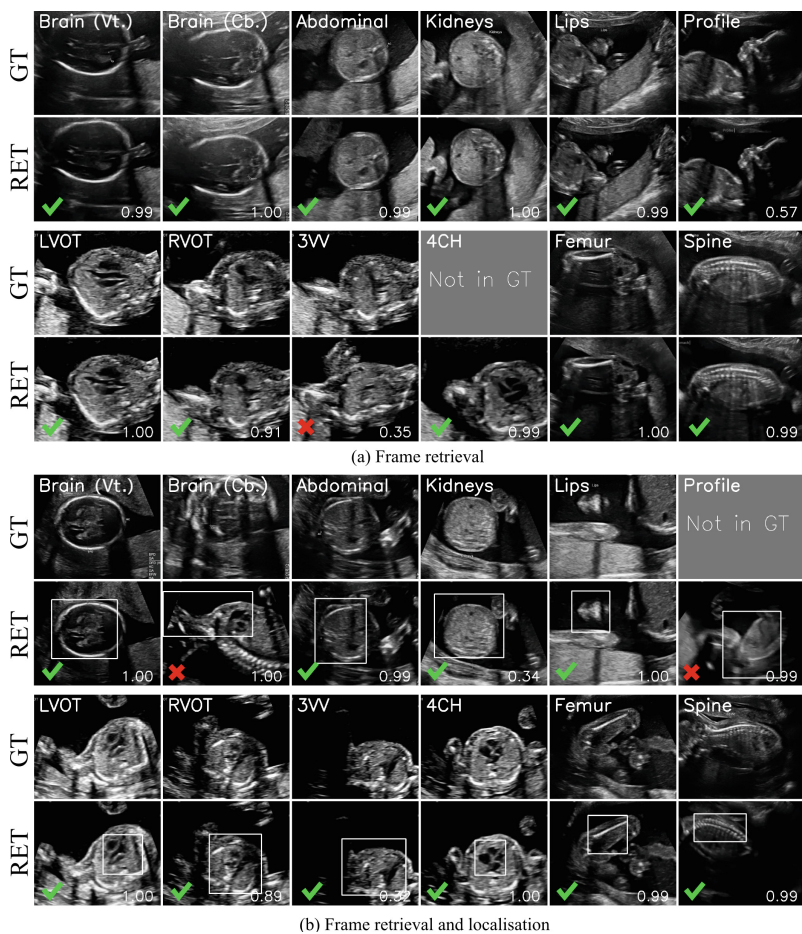


Fig. 4. Retrieved standard frames (*RET*) and *GT* frames annotated and saved by expert sonographers for two volunteers. Correctly retrieved and incorrectly retrieved frames are denoted with a green check mark or red cross, respectively. Frames with no *GT* annotation are indicated. The confidence is shown in the lower right of each image. The frames in (b) additionally contain the results of our proposed localisation (boxes).

our problem is significantly more complex. Using a Nvidia Tesla K80 graphics processing unit (GPU) we were able to classify 113 frames per second (FPS) on average, which significantly exceeds the recording rate of the ultrasound machine of 25 FPS. We include an annotated video in the supplementary material.

Retrospective Frame Retrieval: We retrieved the standard views from videos of all test subjects and manually evaluated whether the retrieved frames corresponded to the annotated *GT* frames for each category. Several cases did not have *GT*s for all views because they were not manually included by the sonogra-

pher in the original scan. For those cases we did not evaluate the retrieved frame. The results are summarised in Table 2. We show examples of the retrieved frames for two volunteers in Fig. 4. Note that in many cases the retrieved planes match the expert GT almost exactly. Moreover, some planes which were not annotated by the experts were nevertheless found correctly. As before, most cardiac views achieved lower scores compared to other views.

Localisation: We show results for the approximate localisation of the respective fetal anatomy in the retrieved frames for one representative case in Fig. 4b and in the supplemental video. We found that performing the localisation reduced the frame rate to 39 FPS on average.

4 Discussion and Conclusion

We have proposed a system for the automatic detection of twelve fetal standard scanplanes from real clinical fetal US scans. The employed fully CNN architecture allowed for robust real-time inference. Furthermore, we have proposed a novel method to obtain localised saliency maps by combining the information in category-specific feature maps with a guided back-propagation step. To the best of our knowledge, our approach is the first to model a large number of fetal standard views from a substantial population of free-hand US scans. We have shown that the method can be used to robustly annotate US data with classification scores exceeding values reported in related work for some standard planes, but in a much more challenging scenario. A system based on our approach could potentially be used to assist or train inexperienced sonographers. We have also shown how the framework can be used to retrieve standard scan planes retrospectively. In this manner, relevant key frames could be extracted from a video acquired by an inexperienced operator and sent for further analysis to an expert. We have also demonstrated how the proposed localised saliency maps can be used to extract an approximate bounding box of the fetal anatomy. This is an important stepping stone for further, more specialised image processing.

Acknowledgments. Supported by the Wellcome Trust IEH Award [102431].

References

1. Bull, C., et al.: Current and potential impact of fetal diagnosis on prevalence and spectrum of serious congenital heart disease at term in the UK. *The Lancet* **354**(9186), 1242–1247 (1999)
2. Chen, H., Dou, Q., Ni, D., Cheng, J.-Z., Qin, J., Li, S., Heng, P.-A.: Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015*. LNCS, vol. 9349, pp. 507–514. Springer, Heidelberg (2015). doi:[10.1007/978-3-319-24553-9_62](https://doi.org/10.1007/978-3-319-24553-9_62)
3. Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P.: Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE J. Biomed. Health Inf.* **19**(5), 1627–1636 (2015)

4. Kurinczuk, J., Hollowell, J., Boyd, P., Oakley, L., Brocklehurst, P., Gray, R.: The contribution of congenital anomalies to infant mortality. University of Oxford, National Perinatal Epidemiology Unit (2010)
5. Lin, M., Chen, Q., Yan, S.: Network in network [arXiv:1312.4400](#) (2013)
6. Maraci, M.A., Napolitano, R., Papageorghiou, A., Noble, J.A.: Searching for structures of interest in an ultrasound video sequence. In: Wu, G., Zhang, D., Zhou, L. (eds.) *MLMI 2014*. LNCS, vol. 8679, pp. 133–140. Springer, Heidelberg (2014). doi:[10.1007/978-3-319-10581-9_17](#)
7. NHS Screening Programmes: Fetal Anomalie Screen Programme Handbook, pp. 28–35 (2015)
8. Ni, D., Yang, X., Chen, X., Chin, C.T., Chen, S., Heng, P.A., Li, S., Qin, J., Wang, T.: Standard plane localization in ultrasound by radial component model and selective search. *Ultrasound Med. Biol.* **40**(11), 2728–2742 (2014)
9. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-Weakly-supervised learning with convolutional neural networks. In: *IEEE Proceedings of CVPR*, pp. 685–694 (2015)
10. Otsu, N.: A threshold selection method from gray-level histograms. *Automatica* **11**(285–296), 23–27 (1975)
11. Salomon, L., Alfirevic, Z., Berghella, V., Bilardo, C., Leung, K.Y., Malinge, G., Munoz, H., et al.: Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound Obst. Gyn.* **37**(1), 116–126 (2011)
12. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net [arXiv:1412.6806](#) (2014)
13. Yaqub, M., Kelly, B., Papageorghiou, A.T., Noble, J.A.: Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015*. LNCS, vol. 9351, pp. 687–694. Springer, Heidelberg (2015). doi:[10.1007/978-3-319-24574-4_82](#)