Automatic and Efficient Standard Plane Recognition in Fetal Ultrasound Images via Multi-scale Dense Networks

Peiyao Kong[1], Dong Ni[1], Siping Chen[1], Shengli Li[2],
Tianfu Wang[1](✉), and Baiying Lei[1](✉)

[1] National-Regional Key Technology Engineering Laboratory for Medical
Ultrasound, Guangdong Key Laboratory for Biomedical Measurement
and Ultrasound Imaging, School of Biomedical Engineering,
Shenzhen University, Shenzhen, China
{tfwang, leiby}@szu.edu.cn

[2] Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare
Hospital of Nanfang Medical University, Shenzhen, People's Republic of China

Abstract. The determination and interpretation of fetal standard planes (FSPs) in ultrasound examinations are the precondition and essential step for prenatal ultrasonography diagnosis. However, identifying multiple standard planes from ultrasound videos is a time-consuming and tedious task since there are only little differences between standard and non-standard planes in the adjacent scan frames. To address this challenge, we propose a general and efficient framework to detect several standard planes from ultrasound scan images or videos automatically. Specifically, a multi-scale dense networks (MSDNet) utilizing the multi-scale architecture and dense connection is exploited, which combines the fine level features from the shallow layers and coarse level features from the deep layers. Moreover, this MSDNet is resource efficient, and the cascade structure can adaptively select lightweight networks when test images are not complicated or computational resources limited. Experimental results based on our self-collected dataset demonstrate that the proposed method achieves a mean average precision (mAP) of 98.15% with half resources and double speeds in FSPs recognition task.

Keywords: Standard plane recognition · Prenatal ultrasound images
Resource efficient · Multi-scale dense networks

1 Introduction

Prenatal diagnosis of fetal abnormalities is quite important for both family and community. 2D ultrasonic examination is the most widely used prenatal diagnostic technique because of its low cost, radiation-free, and the ability to observe the fetus in real time. Prenatal ultrasonography generally involves image scanning, standard planes searching, structural observation, parameter measurement and diagnosis. The determination of standard planes is the precondition of structural observation, parameter measurement and final diagnosis [1], which is a crucial part of antenatal diagnosis.

A. Melbourne and R. Licandro et al. (Eds.): DATRA/PIPPI 2018, LNCS 11076, pp. 160–168, 2018.
https://doi.org/10.1007/978-3-030-00807-9_16

In fact, the judging of the standard plane requires deep knowledge and clinical experience [2]. In the underdeveloped areas, there are lack of the medical resources and experienced doctors. Also, standard plane screening is a time-consuming and laborious task. Therefore, it is of great significance to design an automatic standard plane recognition system, which not only improves the efficiency of prenatal ultrasound examination, but also reduces the burden of doctors.

Due to the continuity of the ultrasound scan images, there are only subtle difference between the standard image and the non-standard image from adjacent frames [3]. Compared with other imaging methods, ultrasound imaging is often affected by noise and artifacts such as shadowing, which results in poor imaging effect and affects the recognition accuracy [4]. As shown in Fig. 1, the first row is the standard plane images, and the second row is the non-standard plane images corresponding to different regions. It can be seen that it is quite difficult for non-professionals to accurately evaluate and distinguish FSPs images. Therefore, recognizing the standard image from ultrasound image automatically is a highly challenging task.

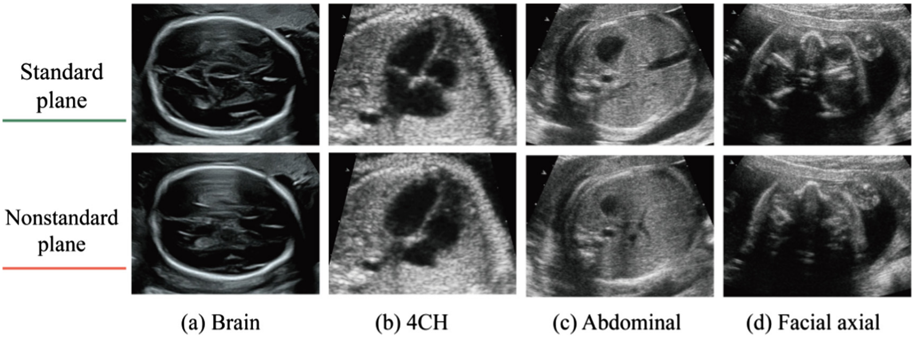


Fig. 1. Illustration of high similarity between standard and non-standard planes in ultrasound images. (a) brain; (b) four channel chamber (4CH); (c) abdominal; (d) facial axial.

In the recent years, deep learning is poised to reshape the feature of machine learning. Over the last decade, research on deep learning has made amazing achievements in many fields. The deep learning related methods has also been widely applied in analyzing medical images for prenatal analysis and diagnosis [5]. In fact, the core concept of deep learning is to learn data representations through increasing abstraction levels, which can learn more abstract and complex representations directly from the raw data. In addition, deep learning has been proved to have stronger applicability and better performance than traditional machine learning methods in the complex image recognition tasks [6]. For this reason, we mainly focus on deep network and representation in this study.

In order to ensure the portability of the algorithm and meet the diagnostic requirements in speed, our study focuses on the resource efficient planning model architecture. However, the previous deep learning studies on the standard image recognition task ignores the computing resources issue when designing the model, which makes the recognition quite slow [7]. Meanwhile, densenet has demonstrated the effectiveness of dense connections in the feature learning process in the related studies since its inception in 2017. For example, Huang et al. built a cascaded network MSDNet [8] based on the idea of dense connections and achieved good classification effect on the CIFAR dataset. Inspired by this, we exploit the MSDNet to build the FSPs recognition architecture. Experimental results on our collected in-house dataset show that our method is easier to mitigate the practical applications to achieve the real time detection in the clinical diagnosis.

2 Methodology

Figure 2 shows the architecture of our proposed method. There are four layers of our network. The specific model design of dense connection and cascade are described in the following sections.

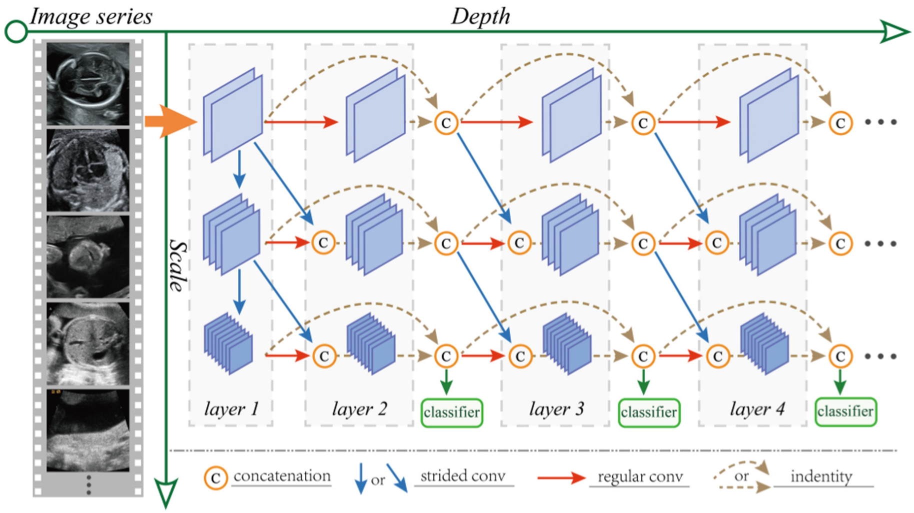
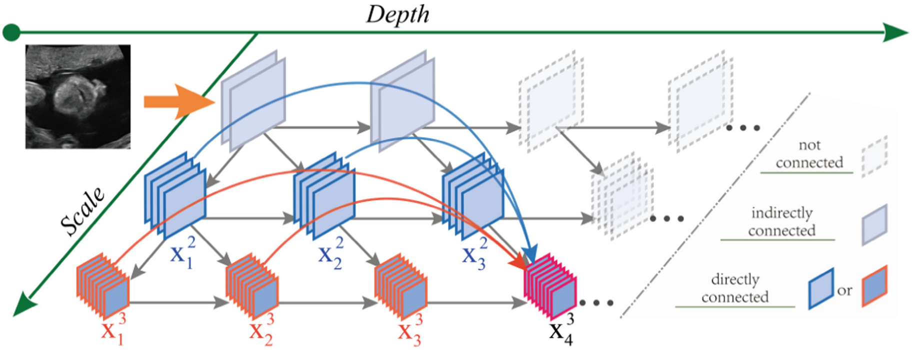


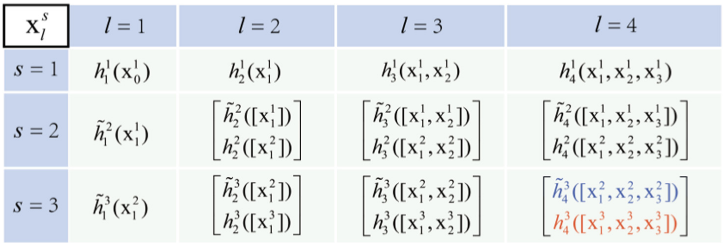
Fig. 2. The first four layers of our network. The horizontal coordinate represents the depth of the network, and the vertical coordinate represents the scale of the feature map. The dense connections across more than one layer are not explicitly drawn: they are implicit through recursive connections.

2.1 Network Architecture

The overall structure of the network is illustrated in Fig. 2. We use Fig. 3 to specify the dense connections in the model. The dense connection mode makes full use of the features with the low-complexity in the shallow layers, which allows the network to reuse and bypass the existing features of the previous layer and ensure high accuracy in later layers [9]. Moreover, dense connections also avoid gradient disappearance, which makes training faster and has less computational power for the same performance.

The network is designed as a cascade of layers that can be split or superimposed depending on the difficulty of different tasks. As can be seen from Fig. 2, there is a classifier designed between each layer and the second layer. This is designed for resource efficient, which enables the model output classification results at any layer of the network. This network adaptively chooses the deeper network for tough task and the shallow network for easy task. The performance of a classifier is located in the shallow layers of a general network, which is often poor due to the lack of coarse scale features. The multi-scale design in the architecture provides coarse scale and high-level feature representations that are amenable to classification.





x_l^s	$l = 1$	$l = 2$	$l = 3$	$l = 4$
$s = 1$	$h_1^1(x_0^1)$	$h_2^1(x_1^1)$	$h_3^1(x_1^1, x_2^1)$	$h_4^1(x_1^1, x_2^1, x_3^1)$
$s = 2$	$\tilde{h}_1^2(x_1^1)$	$\begin{bmatrix} \tilde{h}_2^2([x_1^1]) \\ h_2^2([x_1^2]) \end{bmatrix}$	$\begin{bmatrix} \tilde{h}_3^2([x_1^1, x_2^1]) \\ h_3^2([x_1^2, x_2^2]) \end{bmatrix}$	$\begin{bmatrix} \tilde{h}_4^2([x_1^1, x_2^1, x_3^1]) \\ h_4^2([x_1^2, x_2^2, x_3^2]) \end{bmatrix}$
$s = 3$	$\tilde{h}_1^3(x_1^2)$	$\begin{bmatrix} \tilde{h}_2^3([x_1^2]) \\ h_2^3([x_1^3]) \end{bmatrix}$	$\begin{bmatrix} \tilde{h}_3^3([x_1^2, x_2^2]) \\ h_3^3([x_1^3, x_2^3]) \end{bmatrix}$	$\begin{bmatrix} \tilde{h}_4^3([x_1^2, x_2^2, x_3^2]) \\ h_4^3([x_1^3, x_2^3, x_3^3]) \end{bmatrix}$

Fig. 3. Illustration of dense connections (e.g. x_4^3) and the list of output x_l^s of layer l in scale s.

The vertical connection on the first layer is designed to produce representations on all S scales. It can be thought as an S-layers convolutional network. As shown in Fig. 3, we use x_l^s to represent the output feature maps at layer l and scale s, and the original input is represented as x_0^1. Feature maps at coarser scales are obtained using the down-sampling method.

The feature maps x_l^s of each subsequent layers are a concatenation of all previous feature maps of scale s and $s - 1$. At the bottom of Fig. 3, we have listed the formula for x_l^s of the first four layers. Here, we use $[\ldots]$ to represent concatenation operator, $h_l^s(.)$ is regular convolution, and $\tilde{h}_l^s(.)$ is stride convolution.

In order to test performance of any position in the network, a classifier is designed behind each layer. The classifiers use dense connection within coarsest scale S, such as the classifier at layer l uses all features $\left[x_1^s, \ldots, x_l^s\right]$. Afterwards, we identify the number of layers that are most suitable for our FSPs recognition task by relevant experiments about testing at any location of the model.

For all classifiers, we use cross entropy $L(f_k)$ as a loss function in training. The total cumulative loss functions is defined as

$$\mathcal{L}_{MSD} = \frac{1}{|\mathcal{D}|} \sum\nolimits_{(x,y)\epsilon\mathcal{D}} \sum\nolimits_{k} w_k L(f_k) \tag{1}$$

where $\mathcal{D}$ represents the distribution of training dataset, w_k denotes the weight of the k-th classifier. Empirically, we find that using the same weight for all loss functions works well in practice. In this study, we empirically set the same weight for all the loss functions in our task.

2.2 Data Processing

Our dataset came from acquires 1499 ultrasound examinations of pregnant women with fetal gestational aged from 14 to 28 weeks. All of the data (including images and videos) is compiled from the electronic medical records of the hospital's ultrasound workstation. To some extent, those raw data in the workstation is somewhat cluttered. Unlike some previous studies [10, 11], data is limited to one type of ultrasound device. Our data contains images collected from several brand models of ultrasonic devices consist of Siemens, Samsung, GE, mindray, etc. In order to be more consistent with the actual data distribution, we did not select the data in particular. Therefore, the gap of imaging styles between different devices will be a big challenge for classification and recognition. And then during normal exam, sonographers are used to keep only important standard plane images. Hence all the image data stored in the workstation is basically standard plane. We can only get the non-standard planes from the video set. And sonographers often add pseudo-color to the ultrasound images for more careful observation in some cases. For majority of cases we don't have screen capture videos of entire fetal exam. Only a small number of medical records have short video fragments that record views adjacent to the standard planes. The same as the image data, the short videos also come from multiple branded devices. Each video was acquired from one patient and contained 17–48 frames. We used macro command to extract all their frames.

Because in this study we are only interested in structural information, we removed all color doppler ultrasound images by referring to the practice of ultrasound image data processing in other people's studies [10]. The images contains of doctor's marking and the split screen images showing multiple sections also be removed. In addition all pseudo-color images are converted to grayscale. According to the data situation, we combined our previous work and finally selected six standard plane on the advice of doctors. Finally, we have 22715 ultrasound images in our data set for FSPs recognition task. The detailed composition of data set is shown in Table 1. Moreover we divided the data into training set and test set in a ratio of 4:1.

Table 1. Data summary

Standard planes	Intro	ImageNum
Brain	Horizontal cross section of the thalamus	1840
4CH	Four-chamber view	2409
Abdominal	Standard abdominal view at stomach level	1687
Facial axial	Axial facial view at eyeball level	1585
Facial coronal	Coronal facial view of lips and nose	1959
Facial sagittal	Facial median sagittal view	1725
Others	Unmentioned standard views and non-standard planes	11510

3 Experimental Setting and Results

We implemented all of our models using PyTorch deep learning framework. The training was performed on a single Nvidia GTX Titan Xp, and 64G of RAM. In order to find out the best network depth (l) for our task. We firstly conducted the experiment of five-fold cross validation for different l. Therefore, we randomly divide all data into two parts in a ratio of 1:4, where the small part is used as the final test set and the large part is used for cross-validation. Afterwards, we set the network depth (l) to 15. The results of different depths is collected, and the train epochs is set as 300. We obtain the result of 5 verifications in each classifier, and the average accuracy of five tasks is shown in Fig. 4. In the broken line graph, it can be seen that the recognition accuracy has a significant upward trend at the beginning with the increase of network depth, and it becomes flattens out after the 7th layer. The broken line peaks at the tenth floor, then drops slightly and finally tends to be stable.

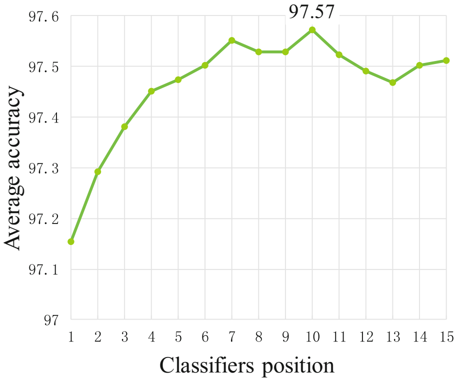


Fig. 4. The average accuracy of five cross validation by classifiers in different depth.

Table 2. Performance comparison of different networks

Model	FLOPs	ACC (%)	FPS
ResNet110	250.81 M	97.23	128.0
DenseNet100	292.23 M	97.64	137.7
Our (*l* = *10*)	**148.01 M**	**98.26**	**226.9**

Based on the verification results in the previous step, we finally set the network depth as 10, take all the data used in the validation as the training set, training epochs is also set as 300. Table 2 shows the comparison of the computation amount, accuracy, and FSPs recognition speed (using frame rate measurements) of different networks. It can be seen that our model achieves nearly twice the speed and half the calculation compared with other networks. Our model obtains a recognition accuracy of 98.25%, which is the highest among all the listed models.

Considering that ‘Others’ class occupies a large proportion in the dataset compared with other classes, we measure the model performance using precision, recall and F1-score for each category. Table 3 shows the detailed test scores for all the standard planes. We can see that our method has achieved good performance in each category, and the average value of all three indicators is over 98%. In addition, the confusion matrix for this test is shown in Fig. 5. From the confusion matrix, we can observe the misclassification occurs between the standard surfaces and ‘others’ class because completely separating standard and non-standard planes is really a hard task.

Table 3. Recognition result (%)

Standard planes	Precision	Recall	F1-score	Images
4CH	98.76	99.38	99.07	481
Abdominal	96.41	95.55	95.98	337
FA	94.74	96.84	95.77	316
FC	98.42	95.40	96.88	391
Brain	100	100	100	368
FS	97.89	94.19	96.00	344
Others	98.45	99.22	98.83	2302
Avg/total	98.15	98.15	98.14	4539

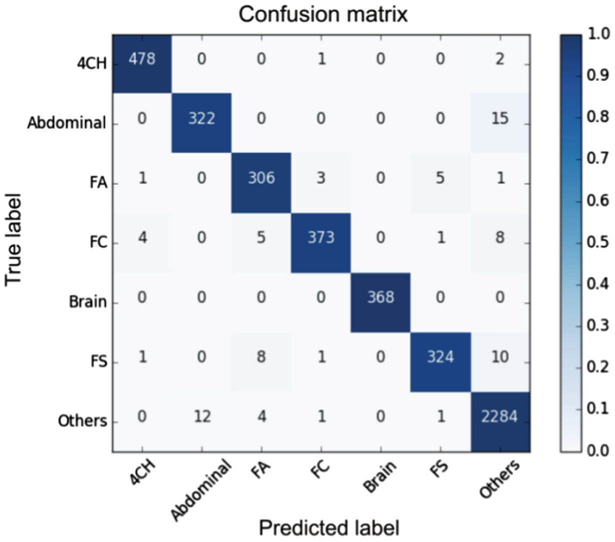


Fig. 5. Confusion matrix for MSD model

For the deep learning model, feature representation has a great impact on the recognition results. In order to more directly demonstrate the effectiveness of our network for FSPs recognition task, we use the t-SNE method [12] to visualize the test data and network feature maps. Specifically, for the original image, we convert the pixels of each image into a row vector and concatenate the values of all the sample vectors along the column dimension. We enter the pixel matrix and their labels into the t-SNE function. Similarly, the output feature vectors before linear layer of classifier are extracted, and t-SNE visualization is performed using the obtained representation form. The visualized results are illustrated in Fig. 6, where different colors in the diagram are used to represent data from different labels. One point in the figure represents one image sample, where a significantly larger number of purple marks represent ‘others’ classes. The left side of the figure is the distribution of the raw data in our testset, and the right side is the data distribution of the network classifier input feature maps (take $l = 10$ as an example). The mixed distribution of test data in the original domain shows that the class differences between FSPs and non-FSPs are very small, which makes our task challenging. We can clearly see that the deep representation after network processing makes the samples have obvious separability, which proves that the proposed model is very effective for FSPs recognition tasks.

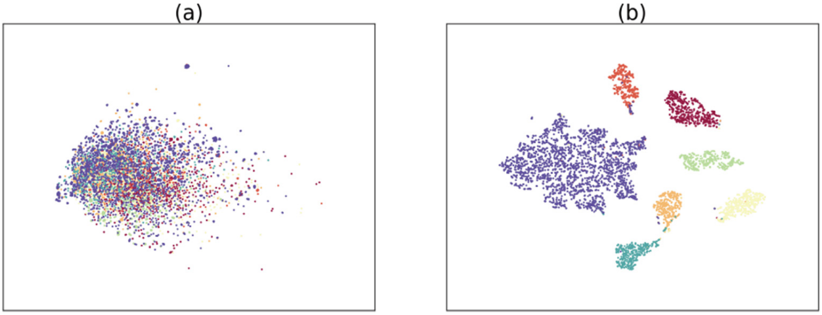


Fig. 6. t-SNE visualization results to illustrate the separability of deep representations in our model. (a) The raw test data distribution; (b) the distribution of data using our network.

4 Conclusion

In this paper, we propose an automatic and efficient FSPs recognition method based on MSDNet with powerful feature representation and efficient cascade design. We verify the effectiveness of our model on the ultrasound standard plane dataset for FSPs recognition task. We obtain the optimal number of network layers for our task through five-fold cross validation. Compared with other networks, the experimental results show that the proposed model achieves quite impressive performance (double speed and half calculations). Finally, through the analysis of multiple indicators, it is proved that our method achieves amazing performance in the recognition of each category. Furthermore, our approach is a general framework and can be extended to the other ultrasound standard planes recognition task. In future work, we will increase the variety of standard planes in the dataset and demonstrate the generalization ability of our model. Also, we will apply this algorithm to real-time detection in clinical practice.

References

1. Li, J., et al.: Automatic fetal head circumference measurement in ultrasound using random forest and fast ellipse fitting. IEEE J. Biomed. Health Inform. **22**, 215–223 (2018)
2. Cai, Y., Sharma, H., Chatelain, P., Noble, J.: SonoEyeNet: standardized fetal ultrasound plane detection informed by eye tracking. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1475–1478. IEEE (2018)
3. Chen, H., et al.: 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, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_62
4. Milletari, F., et al.: Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. Comput. Vis. Image Underst. **164**, 92–102 (2017)
5. Shen, D., Wu, G., Suk, H.-I.: Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. **19**, 221–248 (2017)
6. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
7. Wu, L., Cheng, J.-Z., Li, S., Lei, B., Wang, T., Ni, D.: FUIQA: fetal ultrasound image quality assessment with deep convolutional networks. IEEE Trans. Cybern. **47**, 1336–1349 (2017)
8. Huang, G., Chen, D., Li, T., Wu, F., van der Maaten, L., Weinberger, K.: Multi-scale dense networks for resource efficient image classification. In: International Conference on Learning Representations (2018)
9. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 3 (2017)
10. Baumgartner, C.F., et al.: SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans. Med. Imaging **36**, 2204–2215 (2017)
11. Yu, Z., et al.: A deep convolutional neural network-based framework for automatic fetal facial standard plane recognition. IEEE J. Biomed. Health Inform. **22**, 874–885 (2018)
12. Donahue, J., et al.: DeCAF: a deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Learning, pp. 647–655 (2014)