FETAL FACIAL STANDARD PLANE RECOGNITION VIA VERY DEEP CONVOLUTIONAL NETWORKS

Zhen Yu[1], Dong Ni[1], Siping Chen[1], Shengli Li[2], Tianfu Wang[1*],and Baiying Lei[1*]

Abstract——**The accurate recognition of fetal facial standard plane (FFSP) (i.e., axial, coronal and sagittal plane) from ultrasound (US) images is quite essential for routine US examination. Since the labor-intensive and subjective measurement is too time-consuming and unreliable, the development of the automatic FFSP recognition method is highly desirable. Different from the previous methods, we leverage a general framework to recognize the FFSP from US images automatically. Specifically, instead of using the previous hand-crafted visual features, we utilize the recent developed deep learning approach via very deep convolutional networks (DCNN) architecture to represent fine-grained details of US image. Also, very small (3×3) convolution filters are adopted to improve the performance. The evaluation of our FFSP dataset shows the superiority of our method over the previous studies and achieves the state-of-the-art FFSP recognition results.**

I. INTRODUCTION

The ultrasound (US) image has become a popular way to diagnose the fetal disease accurately in the routine examination [1-6]. The acquisition of the fetal facial standard plane (FFSP) is of vital importance for the accurate diagnosis and measurement [1-3]. Clinically, thorough knowledge and substantial training are required for this task. In US imaging, experienced clinicians can handle with US diagnosis very effectively, but imaging experts and advanced imaging equipment is quite scarce in the under-privileged country. Automatic FFSP recognition from 2-D US images is able to reduce the training time and expedites the physician trial [5, 7]. It is worthwhile and beneficial to develop an automatic diagnosis technology to assist non-experts. The typical way to recognize FFSP is the subjective measurement using the acquired fetal US image.

[1]School of Biomedical Engineering, Shenzhen University, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Shenzhen, China.
[2]Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare Hospital of Nanfang Medical University, 3012 Fuqiang Rd, Shenzhen, P.R.China. (Email:{ tfwang，leibnidong*,}@szu.edu.cn)

However, collecting numerous data and labeling are too time-consuming and impractical in the clinical trial. FFSP often has high within and between class variations caused by a myriad of artifacts such as speckle noises and shadows. Hence, it is a challenging task to recognize different planes. Also, there are no distinguishing difference between FFSP and other planes. To address these challenges, various methods have been proposed in many literatures in the recent years. One of the most common ways is to use the low-level hand-crafted features (i.e., SIFT [1, 3], Haar and HoG) as image descriptor to represent the images. The low-level feature is further encoded by typical method such as bag of visual words (BoVW), vector of locally aggregated descriptor (VLAD), and Fisher vector (FV) [3, 6] to improve the effectiveness of recognition. These methods have been explored to identify the standard plane from US data in the literature as well. However, the existing handcrafted features extracted from consecutive 2D US images are still unappealing for FFSP recognition.

Meanwhile, another trend has been witnessed due to the large available dataset, high performance computing systems (e.g., GPU) and large-scale distributed clusters.[2, 8, 9]. The development of deep convolutional networks (DCNN) [8] has achieved remarkable success in a myriad of fields such as image detection, classification, recognition, segmentation and prediction. Motivated by its powerful representation ability, we adopt a novel DCNN architecture in a bid to obtain the state-of-the-art accuracy of recognizing the standard planes. By exploring the informative feature with great depth, we believe that a very deep CNN architecture is able to boost the recognition performance. Different from the previous methods, a smaller receptive window size and smaller stride of the first convolutional layer than the previous one is adopted.

In this study, we extend our preliminary FFSP recognition work [1, 3] using the deep represented features to enhance performance substantially. We effectively solve the FFSP challenging issue via very deep DCNN architecture. To the best of our knowledge, this is the first automatic FFSP recognition method using very deep CNN represented feature, which has great potential in the practical application in the routine US examination and the prenatal care too.

978-1-4577-0220-4/16/$31.00 ©2016 IEEE

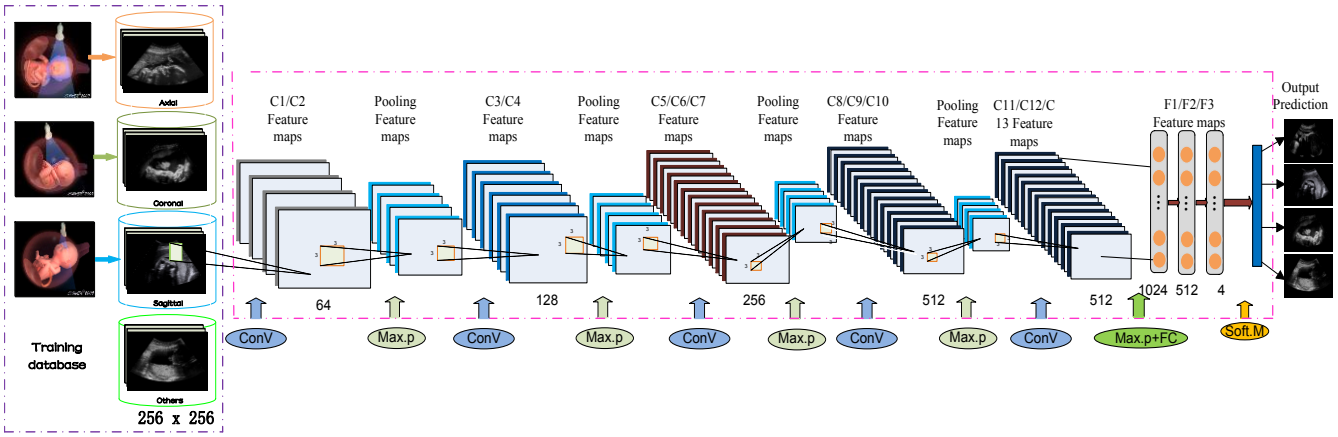

Fig.1. Illustration of our method and DCNN.

II. METHODOLOGY

A. Overview and system architecture

It is known that very deep architecture has achieved state-of-the-art recognition performance in numerous applications [2, 8, 10]. The depth of the CNN is of great importance since it will affect the recognition performance significantly. Among various deep learning method with different lengths, we adopt the same architecture as [8] for our task, which is motivated from its effectiveness and impressive recognition performance. Namely, the entire DCNN layer is configured with the same architecture as [8].

Fig. 1 illustrates the DCNN architecture and deep representation of the US images. The input layer is 256×256 gray-scale US images. The network is composed of 13 convolutional (Conv.) layers, 5 max pooling layers and 3 fully connected (FC) layers. All the kernels size are 3x3. A total of 5 spatial pooling is performed using 5 max-pooling layers. In DCNN configurations for FFSP recognition, we follow the VGG-16 weight layers design in the network (8 conv. and 3 FC layers) [8]. Noted that the US images is first preprocessed (i.e., script removal, image enhancement and noise reduction, subtract the mean value that computed only over the training data) to enhance the recognition performance. After preprocessing, the region containing axial, coronal and sagittal plane (namely region of interest, ROI) is fed into the DCNN architecture.

The large filters can be stacked by small filter, which is able to reduce the parameter numbers and increase the network nonlinearity. Hence, we obtain our compact and discriminative feature representation using small filters. Specifically, we employ a smaller 3×3 receptive field to get the information of left/right, up/down, center in a stack of convolutional field rather than a larger receptive field. The reason is that a stack of two 3×3 conv. layers has an effective receptive field of 5×5, and three such layers have a 7×7 effective receptive field [8]. The convolution stride and pad are both set to 1 pixel. A 2×2 pixel window with stride 2 is applied through the max-pooling. We utilized a fixed-size image scale in each pixel. All hidden layers are equipped with the rectification non-linearity (ReLU). Similarly, the channels of feature maps of Conv. layers increase from 64 to 512 (i.e., the depth dimension of activation maps) in our networks.

The FC layers contain three layers with 1024-512-4 channels, and one classification layer with 4 channels (one for each class). The soft-max layer is the last layer. Despite the large depth, the number of weights only changes slightly. We also use a regularization scheme to decrease the weight and avoid the overfitting. The regularization is achieved by dropout for the first two fully-connected layers due to the large number of parameters. We empirically set it to 0.5 to be discriminative and generalized to separate the planes.

It is known that bad initialization will stop learning because of the gradient instability in deep nets, and hence the appropriate initialization is quite important. To address this issue, we start with very deep DCNN configuration for training task with pre-training from ImageNet dataset for initialization. Namely, our FFSP training is based on the pre-trained model from ImageNet dataset with the deep architectures. The learning rate is decreased from 0.001 steadily.

B. FFSP recognition via deep Convolutional Networks

Let $x_c^i(c = 1, \dots, C)$ be the input image with respect to *c*-th class. $l_c^i \in \{0,1\}$ is the corresponding ground-truth labels for the input FFSP image x_c^i. Let $\boldsymbol{l}^i = [l_1^i, l_2^i, \dots, l_C^i]$ be the label vector for the *i*-th image. If the US image is correctly predicted as class *c*, l_c^i=1, otherwise, l_c^i=0. The probability to predict the ground truth correctly is defined as $\boldsymbol{l}_i / ||\boldsymbol{l}_i||_1$. Our convolution layer is denoted as:

$$y^{j(r)} = \max\left(0, \sum_i k_c^{ij(r)} * x_c^{i(r)} + b_c^{j(r)}\right) \qquad (1)$$

where x_c^i and y_c^j are the *i*-th input map and the *j*-th output map for the *c*-th class, respectively. k_c^{ij} is the convolution kernel between the *i*-th input map and the *j*-th output map. $*$ denotes convolution. b_c^j is the bias of the *c*-th class for the *j*-

th output map. r means the shared weights for the local regions (i.e., we locally shared in every 3×3 regions in the first layer). Due to the better fitting performance than sigmoid function, we use the ReLU nonlinearity for hidden neurons as below:

$$R(x) = \max(0, x), \quad (2)$$

where $R(x)$ is the ReLu activation function. In our case, the weights in higher convolutional layers of the DCNN architecture are locally shared to learn different middle or high level features in different regions.

The max-pooling is computed by:

$$y_{j,k}^{i} = \max_{0 \le m,n < s} \{x_{j.s+m,k.s+n}^{i}\}, \quad (3)$$

where each neuron in the i-th output map y^i pools over an $s \times s$ non-overlapping local region in the i-th input map x^i. The last hidden layer takes the function:

$$y_j = \max\left(0, \sum_i x_i^1 . w_{i,j}^1 + \sum_i x_i^2 . w_{i,j}^2 + b_j\right), \quad (4)$$

where x_1 , x_2 , w_1 , w_2 are neurons and weights in each corresponding Conv. layers, respectively. The two Conv. layers integrate features linearly, and ReLU non-linearity is followed after the linear combination. An n-way softmax is the output predictions using the probability distribution over n different identities:

$$y_i = \frac{\exp(y_i')}{\sum_{j=1}^{n} \exp(y_j')}, \quad (5)$$

where $y_j' = \sum_{i=1}^{f} x_i . w_{i,j} + b_j$ linearly combines a total of f features, x^i as the input of neuron j, and y_j' is its output. Accordingly, DCNN predicts c-th target class by minimizing the cost function. We employ stochastic gradient descent to optimize it via the gradients calculated from back propagation.

To classify numerous FFSP US images in the training data, the low-dimensional feature should contain strong discriminative information from all the US fetal images. Similarly, given a testing FFSP US image, it is classified into the target FFSP US image using the classifier layer trained by DCNN architecture. The class score for output predictions is a fixed-size vector, which is obtained by the spatially average (sum pooling). That is, the final scores for the FFSP US images are obtained by averaging the class posteriors from the soft-max output. Hence, this strong and efficient classifier is devised to recognize each plane.

III. EXPERIMENTAL RESULTS

To evaluate FFSP recognition performance, we perform the experiment using our self-collected US images dataset extracted from US videos and acquired by an US scanner from Siemens Acuson Sequoia 512 from Shenzhen Maternal and Child Health Hospital. Our fetal gestational ages range typically from 20 to 36 weeks. An experienced obstetrician annotates the entire US images manually. Our dataset is composed of 187 images of axial plane, 192 images of coronal plane, 203 images of sagittal plane, and 1153 of others randomly extracted from the rest images without any standard plane. Our testing is based on a total of 166 images (14 axial, 27 coronal, 25 sagittal and 99 others image) randomly selected from each class and the rest is for training. FFSP US image size is rescaled from 576×768 to 256×256 for DCNN training. The accuracy, precision; recall, true positive rate (TPR) and false positive rate (FPR) and the corresponding curves are obtained to evaluate the performance.

Our evaluation is based on the publicly available C++ Caffe toolbox. We use GPUs installed in a single system. The gradient of the full batch is performed by averaging each batch. The complicated algorithm is able to speed up the training, but this simple scheme is fast enough to achieve our goal. With equipped NVIDIA Titan Black GPUs, training a single net took 2–3 hours on the architecture.

The confusion matrix of FFSP standard plane recognition is illustrated in Fig.2, where the rows are the actual standard plane, and the columns represent the predicted plane. The ratio of the correctly predicted over the total testing US image is the overall recognition accuracy. The overall recognition plane is 96.99%, which is high enough for the practical application.

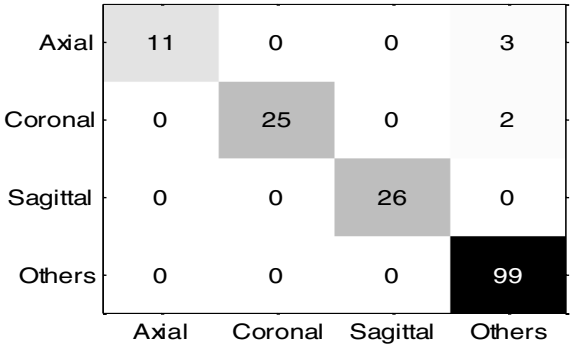

Fig.2. Confusion matrix of FFSP recognition results.

The ROC curves in terms of TPR and FPR are illustrated in Fig.3. It can be seen that FFSP recognition results for these three standard planes achieved promising results. We observe that the axial plane is the most challenging plane to be recognized since the error is the highest among the rest, while sagittal plane is easy to separate as the recognition rate is 1. The recognition performance is consistent with our previous work. The primary explanation is that axial plane is too complicated, which is easy to be confused with others in US imaging. Accordingly, more discriminative power is needed to distinguish it. Fig. 4 shows the precision and recall curves. It can be seen that both ROC and PR curves have the consistent result and similar characteristics.

Apart from the above evaluation, Table 1 gives the detailed FFSP recognition results using DCNN method. It is observed that the highest recognition results are obtained for each plane. Axial and coronal plane are easily confused with other plane since the FP results are high. Table 2 shows the average recognition results of our method and comparison with the typical DSIFT visual feature encoded by the traditional feature encoding algorithm (i.e., BoVW, VLAD, FV and multi-layer FV (MFV) [1, 3]. Although feature encoding is

able to separate different planes with relatively good performance using the low-level hand-crafted visual features, we observe that DCNN obtains the best recognition performance and outperforms the traditional methods.

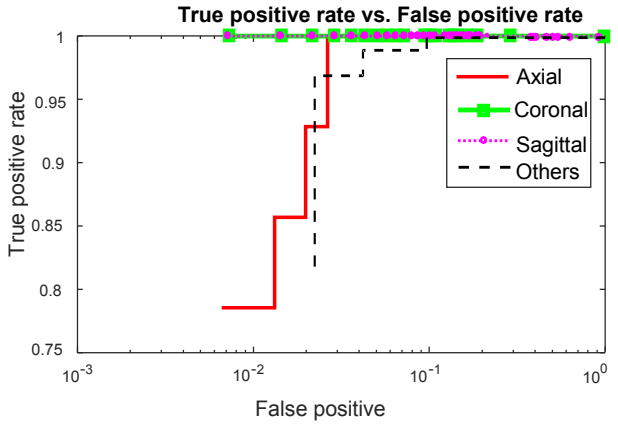

Fig.3. True positive and false positive curve.

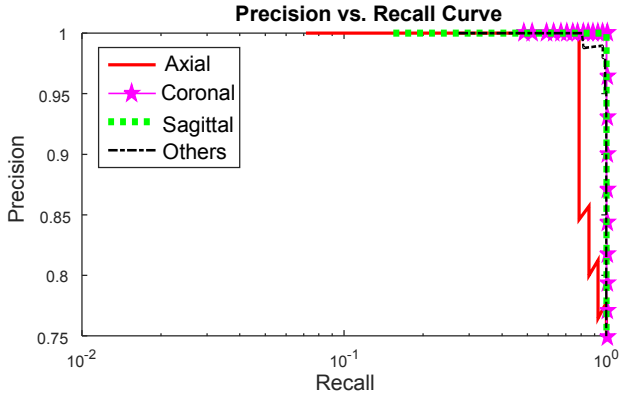

Fig.4. Precision and recall curve.

Table 1. Recognition results in terms of different planes.

Plane	True Positive	False Positive	TPR	FPR	Precision	Recall
Axial	11	3	98.1%	78.6%	99.1%	89.3%
Coronal	25	2	98.6%	92.6%	99.3%	96.3%
Sagittal	26	0	1	1	1	1
Others	99	0	1	96.1%	98.1%	97.0%

Table 2. Algorithm comparison for FFSP recognition (%).

Algorithm	Acc	TPR	FPR	Precision	Recall
DSIFT-BoVW	87.39	85.47	85.68	86.59	86.34
DSIFT-VLAD	89.61	88.95	87.69	87.55	88.14
DSIFT-FV	90.84	91.27	91.78	90.25	91.56
DSIFT-MFV	93.21	93.89	94.56	93.14	93.87
DCNN	96.99	96.98	98.49	96.98	98.99

IV. CONCLUSIONS

In this work, we propose a new FFSP recognition results in US images with a very deep CNN architecture. Instead of using shallow and large convolutional layer, we use a 16-layer weight and a 3×3 small convolution size to further improve the performance. Experimental results demonstrate that DCNN method has achieved promising FFSP recognition results and outperformed the traditional method significantly. Our future work will focus on the evaluation of the effect of depth and the collection of larger FFSP dataset. We also will focus on fusion method, which is able to further enhance the performance.

ACKNOWLEDGEMENT

This work was supported partly by National Natural Science Foundation of China (Nos. 61402296, 61571304, 81571758, 61501305 and 61427806), Shenzhen Key Basic Research Project (Nos. JCYJ20150525092940986, JCYJ2015052509 2940982,JCYJ20130329105033277 and JCYJ20140509172 609164), the (Key) Project of Department of Education of Guangdong Province (No. 2014GKXM052) and Shenzhen-Hong Kong Innovation Circle Funding Program (No.JSE201109150013A).

REFERENCES

[1] B. Lei, L. Zhuo, S. Chen, S. Li, D. Ni, and T. Wang, "Automatic recognition of fetal standard plane in ultrasound image," in Proc. of ISBI, 2014, pp. 85-88.

[2] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P.-A. Heng, "Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks," IEEE Journal of Biomedical and Health Informatics, vol. 19, pp. 1627-1636, 2015.

[3] B. Lei, E.-L. Tan, S. Chen, L. Zhuo, S. Li, D. Ni, and T. Wang, "Automatic Recognition of Fetal Facial Standard Plane in Ultrasound Image via Fisher Vector," PLoS ONE, vol. 10, p. e0121838, 2015.

[4] B. Rahmatullah, A. Papageorghiou, and J. A. Noble, "Automated Selection of Standardized Planes from Ultrasound Volume," in Machine Learning in Medical Imaging. vol. 7009, 2011, pp. 35-42.

[5] L. Zhang, S. Chen, C. T. Chin, T. Wang, and S. Li, "Intelligent scanning: Automated standard plane selection and biometric measurement of early gestational sac in routine ultrasound examination," Medical Physics, vol. 39, pp. 5015-5027, 2012.

[6] B. Lei, Y. Yao, S. Chen, S. Li, W. Li, D. Ni, and T. Wang, "Discriminative Learning for Automatic Staging of Placental Maturity via Multi-layer Fisher Vector," Scientific reports, vol. 5, 2015.

[7] D. Ni, T. Li, X. Yang, J. Qin, S. Li, C.-T. Chin, S. Ouyang, T. Wang, and S. Chen, "Selective Search and Sequential Detection for Standard Plane Localization in Ultrasound," in Abdominal Imaging. Computation and Clinical Applications. vol. 8198, 2013, pp. 203-211.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[9] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in Proc.of CVPR, 2014, pp. 1891-1898.

[10] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang , V. C. Mok, L. Shi, P.-A. Heng, "Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks," IEEE Transactions on Medical Imaging, vol. 35, pp. 1182-1195, 2016.