AUTOMATIC REAL-TIME CNN-BASED NEONATAL BRAIN VENTRICLES SEGMENTATION

Puyang Wang[1], *Nick. G. Cuccolo*[2] *and Rachana Tyagi*[2] *and Ilker Hacihaliloglu*[3,4] *and Vishal M. Patel*[1]

[1] Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ
[2] Department of Surgery, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ
[3] Department of Biomedical Engineering, Rutgers University, Piscataway, NJ
[4] Department of Radiology, Rutger Robert Wood Johnson Medical School, New Brunswick, NJ, USA

ABSTRACT

Quantitative imaging of brain plays an important role in preterm neonates with a very low birth weight due to the increased risk of developing intraventricular hemorrhage (IVH). In this work, we propose a fully automated method for segmentation of ventricles from two-dimensional (2D) ultrasound (US) scans. The proposed method is based on a Convolutional Neural Network (CNN) that combines the advantages of U-Net and SegNet architectures for ventricles segmentation. Extensive experiments on a dataset consisting of 687 US scans show that the proposed method achieves significant improvements over the state-of-the-art medical image segmentation methods.

Index Terms— Segmentation, ultrasound images, ventricles segmentation.

1. INTRODUCTION

Very low birth weight ($< 1,500g$) premature babies account for 1.4 percent of all births in the United States. Of those, more than 16,000 babies each year will develop intraventricular hemorrhage (IVH) [1]. These hemorrhages result in ventricle dilation, which can lead to serious brain damage if not properly treated. Monitoring of ventricle volume change in neonates is clinically important in order to determine the correct intervention. Two-dimensional (2D) ultrasound (US) is currently the main imaging modality used in the diagnosis and monitoring of IVH. However, irregular shape deformation of ventricles, high levels of noise and various imaging artifacts present in the acquired ultrasound data results in the inability to localize the site and extent of brain injury, or to predict neurologic outcomes in identifying IVH from US data. Due to these difficulties, quantitative assessment of anatomical information is mostly performed manually or using semi-automated methods [2]. In [3], a fully automated atlas-based segmentation pipeline was developed for segmenting 3D volumetric US data. Validation results performed on 30 3D US scans achieved a mean Dice similarity coefficient (DSC) and maximum absolute distance of 76.5% and 1 mm, respectively. The reported computation time for segmenting a single 3D volume was 54 mins [3].

In recent years, deep learning-based methods have shown to produce state-of-the-art results in many computer vision and medical image analysis tasks [4], [5], [6]. Inspired by [7], various CNN-based encoder-decoder networks have been proposed for different computer vision tasks in the literature. A typical encoder-decoder structure starts with an encoder network which decreases spatial resolution while learning a high-dimensional representation, followed by a decoder that recovers the original input resolution and outputs low-dimensional predictions. In particular for biomedical image segmentation, one such network, called SegNet was recently proposed in [8] and has been widely used for medical image segmentation. A conditional generative adversarial network-based method called pix2pix was proposed in [9]. U-Net is another popular network that has been widely used in the medical imaging community for segmentation [10]. The advantage of U-net structure comes from the symmetric contracting and expanding path which is capable of leveraging contextual information of different scales.

In this paper, we propose a new CNN encoder-decoder structure that combines the advantages of both SegNet and U-Net. The proposed method features an encoder that extracts deep features and a decoder that leverages multi-scale information contained in the extracted deep features. We validate our proposed network against state-of-the-art segmentation methods. Figure 1 gives an overview of the proposed brain ventricles segmentation network.

2. PROPOSED METHOD

In this section, we provide details of the proposed US image segmentation method in which we aim to learn a mapping function between input US scans and the manual segmentation result using a specially designed CNN. The proposed method consists of two main components: deep feature extractor (i.e. encoder) and multi-scale decoder.

For extracting features, we can use one of the many pre-

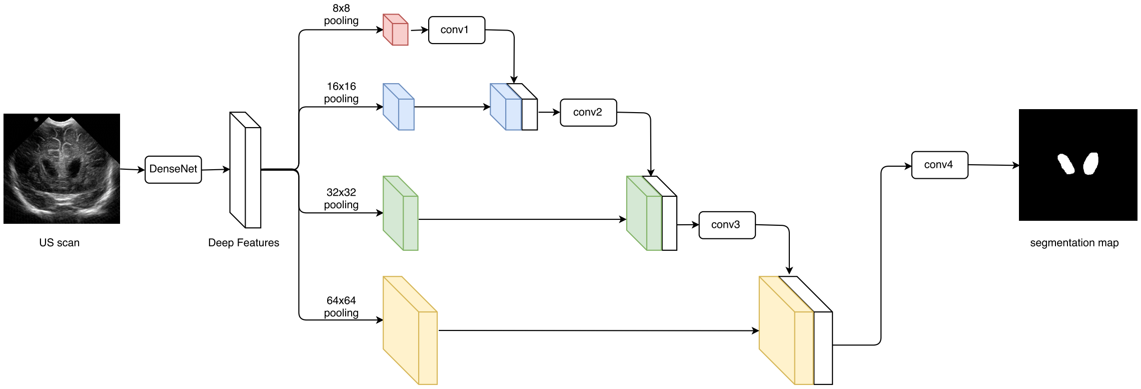

Fig. 1: An overview of the proposed CNN architecture for brain ventricles segmentation from ultrasound images.

trained CNNs proposed in the literature. These CNNs often consist of either deep or shallow networks. However, increasing the depth of a network often results in an optimization difficulty. This problem has been partially solved by ResNets [11] and Highway Networks [12] with skip-connections. Recently, in [13], a novel deep neural network called DenseNet which connects each layer to every other layer in a feed-forward fashion is proposed and shown to outperform both ResNets and Highway Networks in image classification. In the proposed method, we use a pretrained DenseNet as the encoder to take the advantage of very deep neural network.

As for the decoder, the primary goal is to classify each pixel into one of two classes (ventricle or non-ventricle) given the extracted deep features. In other words, the decoder translates the input high-dimensional deep features into binary images (0: non-ventricle, 1: ventricle). As noted in U-Net [10], the key part of precise pixel-wise prediction for biomedical image segmentation task is to make good use of the multi-scale features. In our method, we accomplish this task by first pooling the feature maps into four different sizes followed by a series of transposed convolutions that transform lower dimensional feature maps into higher ones in steps.

The detailed architecture of the proposed method is illustrated in Figure 1, where $conv$ denotes a sequence of transposed convolution, batch normalization and rectified linear unit (CONV-BN-ReLU), respectively. Note that the output of each transposed convolution is concatenated with existing feature maps of the same size and then fed into the next transposed convolution.

2.1. Network Architecture

As shown in Figure 1, we use pooling to downsample the feature maps extracted by the DenseNet into four different sizes: $8 \times 8, 16 \times 16, 32 \times 32$, and 64×64. Each of four pooled feature map has C channels, where C is number of channels in the previous original feature map. This pooling process is similar to the contracting path in U-Net but instead of the max pooling operation, we use adaptive average pooling which enables arbitrary large input size.

In order to generate the output of the same size as the input from all four different sized feature maps, upsampling is necessary. Despite many upsampling techniques, we choose transposed convolution. The decoder network starts with a transposed convolution (3×3 kernel size, stride 2 and padding 1) on the smallest pooled feature map ($C \times 8 \times 8$). As a result of stride 2, the size of the output feature map is doubled. A concatenation of the output and the corresponding pooled feature map of the same size is then fed into next the transposed convolution that has the same kernel size, stride and padding as the first one. Same process is repeated for the remaining pooled feature maps. However, because of concatenation, the number of feature channels reduce by half except for the first convolution.

Finally, in the last $conv$ block as shown in right part of Figure 1, a number of L transposed convolutions transform the feature map ($C \times 64 \times 64$) into the final segmentation map. Here, L depends on the desired output size. The configuration of each convolution layer is given in Table 1. Here, $conv1$ to $conv4$ denote the sequence of Conv-BN-ReLU layers as depicted in Figure 1. Note that the desired final output image size is assumed to be 512×512. Hence, three transposed convolutions ($L = 3$) is needed for $conv4$.

Table 1: Network Configuration.

Block	Layer	Kernel Size	# Filters	Stride	Output Size
DenseNet					$C \times H \times W$
conv1	conv(1)	$C \times 3 \times 3$	C	2	$C \times 16 \times 16$
conv2	conv(2)	$2C \times 3 \times 3$	C	2	$C \times 32 \times 32$
conv3	conv(3)	$2C \times 3 \times 3$	C	2	$C \times 64 \times 64$
conv4	conv(4)	$2C \times 3 \times 3$	C	2	$C \times 128 \times 128$
	conv(5)	$C \times 3 \times 3$	C/2	2	$C/2 \times 256 \times 256$
	conv(6)	$C/2 \times 3 \times 3$	1	2	$1 \times 512 \times 512$

2.2. Training Details

In this section, we provide the details of training our proposed network including dataset, loss function and training parameters.

Data collection: After obtaining the approval from the Rutgers University Institutional Review Board, retrospective brain US scans were collected from subjects who were treated at the Robert Wood Johnson Medical Hospital. De-identification of the data is performed before using them for further processing. A total of 687 in vivo B-mode US images are collected. All the ventricles were manually segmented from the collected scans by an expert. Figure 2 shows a sample image and the corresponding ground truth image from this dataset.

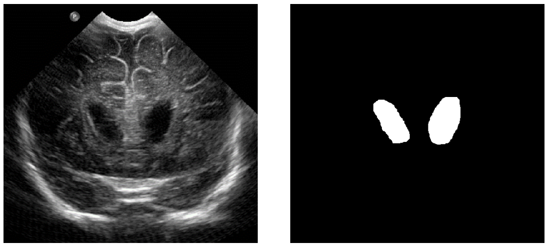
(a) US scan (b) Manual Segmentation

Fig. 2: Sample brain US scan and the corresponding manual ventricles segmentation.

Data augmentation: Since a deep CNN network often requires a large number of training samples, we perform data augmentation to generate extra training samples from the original data. The input data is augmented using horizontal flip and random crop. We perform horizontal flip to every samples, therefore the total number of data is doubled. Furthermore, all images in the dataset are first resized to 600×600 and then randomly cropped to the size of 512×512 during training. By applying this data augmentation strategy, we generate sufficient samples to eliminate as much dataset bias as possible.

Loss function: Given an image and segmented map pair (Y, X), where Y is the input US image and X is the corresponding segmentation ground truth, the per-pixel L1 loss is defined as

$$L(\phi) = \frac{1}{WH} \sum_{w=1}^{W} \sum_{h=1}^{H} \|\phi(Y^{w,h}) - X^{w,h}\|_1, \qquad (1)$$

where ϕ is the learned network (parameters) and X and Y are assumed to have the same size of $W \times H$. By using this loss function, the network is trained to minimize the L1 distance between the output and the ground truth on the training set.

Training parameters: Among many different configurations of DenseNet, we choose the pretrained DenseNet121 as our encoder and re-train it along with the decoder. Since both the input and output images are of size 512×512, the number of transposed convolutions with stride 2 in the last $conv$ block should be set to 3. The entire network is trained using the ADAM optimization method [14], with mini-batches of size 12 and learning rate of 0.0002.

3. EXPERIMENTAL RESULTS

For evaluations, we randomly select 50 samples from the whole dataset of 687 samples as the test set. The remaining 637 samples are used as the training set. The network was trained for 100 epochs to ensure the convergence of the loss function. After the network was trained, we evaluate it on the test set. We compare the performance of our method with that of the following three recent methods: SegNet [8], U-net [10] and pix2pix [9] . For all the compared methods, parameters are set as suggested in their corresponding papers and trained using the same training dataset as used to train our network.

Experiments are carried out three times. The Dice coefficient, Intersection over Union (IoU) and pixel-wise accuracy (Pixel Acc.) are used to measure the segmentation performance of different methods. Average results corresponding to three randomized tests are shown in Table 2. As can be seen from this table, in all three metrics, our method provides the best performance compared to the other methods. This experiment clearly shows the significance of the proposed multi-scale decoder for image segementation.

Table 2: Comparison of the proposed method with SegNet [8], U-Net [10] and pix2pix [9].

Method	DICE	Mean IoU(%)	Pixel Acc.(%)
SegNet [8]	0.876±0.111	80.35±0.178	87.64±0.138
U-Net [10]	0.889±0.080	82.33±0.120	89.52±0.133
pix2pix [9]	0.869±0.103	79.89±0.137	88.64±0.130
Our	**0.908±0.053**	**84.84±0.078**	**92.14±0.063**

Apart from the quantitative comparison, we also compare our method with others qualitatively by visual inspection. The segmentation results corresponding to different methods on two input US scans are shown in Figure 3. The second to the fifth columns of Figure 3 show the segmentation maps corresponding to SegNet, U-Net, pix2pix and our method, respectively. The ground truth segmentation maps are shown in the last column of this figure. It can be observed that quantitative results are consistent with the visual results. No artifacts exist in our method while SegNet and pix2pix suffer from some noticeable artifacts for the first sample. It is also evident from the second row of Figure 3 that our method is capable of segmenting both small and large ventricles reasonably well compared with the other methods. This clearly demonstrates the effectiveness of the proposed multi-scale decoder for US image segmentation.

Experiments were carried out with an Intel Xeon CPU at 3.00GHz and an Nvidia Titan-X GPU with 8GB of memory. On average our method takes about 22ms to segment an US image of size 512×512, which is sufficient for real-time applications.

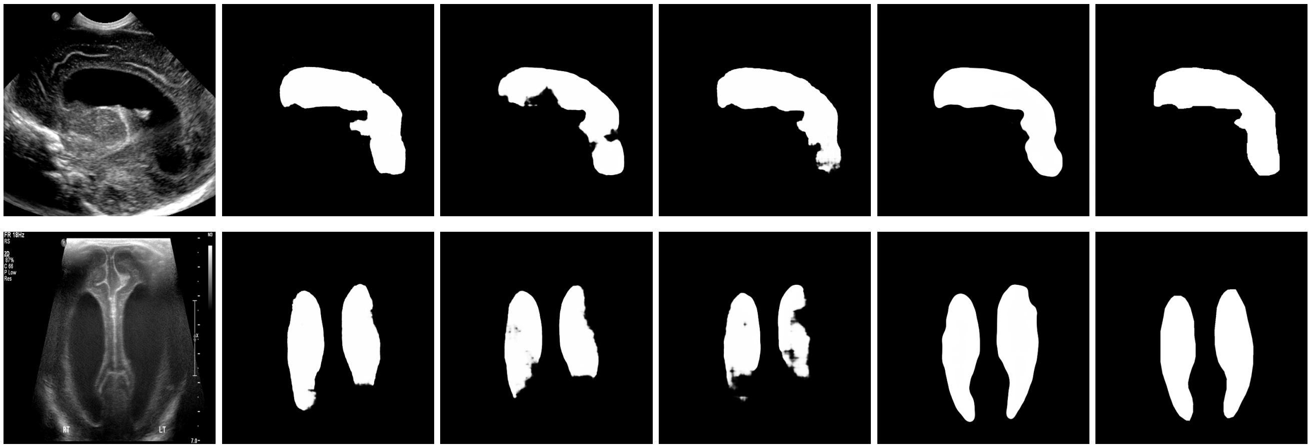

Fig. 3: From left to right: B-mode US scans, SegNet [8], U-Net [10], pix2pix [9], our, manual segmentation.

4. CONCLUSION

The achieved results are promising for further investigation. The proposed CNN architecture achieves improved qualitative and quantitative results over previous state-of-the-art. The reported real-time computational time makes the method suitable for bedside investigation purposes. To the best of our knowledge, this work reports the first study on fully automatic real-time segmentation of ventricles from 2D US data. Future work will involve validation of the proposed methods on more clinical scans and extension to 3D for processing volumetric US data.

5. REFERENCES

[1] Shenandoah Robinson, "Neonatal posthemorrhagic hydrocephalus from prematurity: pathophysiology and current treatment concepts: a review," *Journal of Neurosurgery: Pediatrics*, vol. 9, no. 3, pp. 242–258, 2012.

[2] Wu Qiu, Jing Yuan, Jessica Kishimoto, Jonathan McLeod, Yimin Chen, Sandrine de Ribaupierre, and Aaron Fenster, "User-guided segmentation of preterm neonate ventricular system from 3-d ultrasound images using convex optimization," *Ultrasound in medicine & biology*, vol. 41, no. 2, pp. 542–556, 2015.

[3] Wu Qiu, Yimin Chen, Jessica Kishimoto, Sandrine de Ribaupierre, Bernard Chiu, Aaron Fenster, and Jing Yuan, "Automatic segmentation approach to extracting neonatal cerebral ventricles from 3d ultrasound images," *Medical image analysis*, vol. 35, pp. 181–191, 2017.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[5] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] Hayit Greenspan, Bram van Ginneken, and Ronald M Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[7] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[8] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[13] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[14] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.