Segmentation of neonates cerebral ventricles with 2D CNN in 3D US data: suitable training-set size and data augmentation strategies

Matthieu Martin
CREATIS Laboratory, Lyon, France*
matthieu.martin@creatis.insa-lyon.fr

Bruno Sciolla
CREATIS Laboratory, Lyon, France*

Michaël Sdika
CREATIS Laboratory, Lyon, France*

Philippe Quétin
CH Avignon, Avignon, France

Philippe Delachartre
CREATIS Laboratory, Lyon, France*

*Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69100, LYON, France.

***Abstract*—For its clinical potential, segmentation of the preterm neonate's Cerebral Ventricular System (CVS) in 3D ultrasound (US) data using convolutional neural networks (CNN) is an emerging field. Nevertheless, gathering manually annotated data to efficiently train a CNN is difficult. In this paper, we address the question of how many training volumes and which kind of artificial data augmentation strategies would be suitable for this application. We explored this topic by training a U-net with different training-set size and by using several artificial data augmentation strategies. In our set-up, accurate segmentation results ($Dice \geq 0.8$) were obtained with only 9 training volumes. The use of artificial data augmentation improved significantly ($p < 0.05$) the accuracy obtained without data augmentation when performing horizontal flips (between the right and the left brain hemispheres). The other types of data augmentation that we tried did not significantly improve U-net accuracy.**

***Index Terms*—preterm neonates, CVS 3D segmentation, 3D US data, training-set size, artificial data augmentation**

I. Introduction

Automatic segmentation of the CVS in 3D US data is an emerging field for its clinical potential. Indeed, preterm neonates are highly likely to suffer from ventriculomegaly which corresponds to an abnormal accumulation of cerebrospinal fluid inside the CVS which causes its dilation. The estimation of the CVS volume through its automatic segmentation in 3D US data would enable its monitoring for all preterm neonates. This clinical parameters is needed to decide when a surgical intervention must be performed to reduce intra-cranial pressure and to prognosticate the neurodevelopmental outcome of these patients [1].

Preterm neonates' CVS segmentation in 3D US data was achieved by [2] and [3]t using atlas-based registration followed by active contour segmentation methods and recently by [4] using a 2D U-net.

CNNs have recently outperformed the state-of-the-art methods in image segmentation problems [5], [6]. These methods enable the segmentation of difficult structures, in few seconds, with an outstanding level of accuracy. Thus, they are perfectly suited to this segmentation problem. In particular U-net [7] and V-net [8] are CNNs architecture which have been widely used to solve medical image segmentation problems. Yet, to tackle a segmentation problem with a CNN, it is necessary to build a database. This step is difficult and very time-consuming, especially for a medical image segmentation problem which requires a physician to acquire and annotate 3D data. In addition, the suitable number of training volumes required and the suitable kind of artificial data augmentation to be used are application dependent. To the best of our knowledge, these questions have not been studied for this application.

In this paper, we trained a 2D U-net to achieve the CVS segmentation in 3D US data. We propose to estimate the number of training volumes that are necessary to obtain an accurate precision ($Dice \geq 0.8$) and to determine which kind of artificial data augmentation strategies are suitable for this application. We found that only 9 training volumes were necessary to obtain accurate segmentation results and that performing horizontal flips during the training process could improve the network accuracy.

II. Method

A. Data description

In this study, 25 3D reconstructed US data created with the algorithm we proposed in [4] were used. They were reconstructed from 2D freehand transfontanellar ultrasonography that were acquired by a paediatrician in coronal orientation with a multi-D Siemens Acuson 9L4 probe. The patients mean age at acquisition was 35.81 ± 1.59 weeks of amenorrhea. The data size was $320 \times 320 \times 320$ and all the 3D US volumes contained a 3D segmentation of the CVS. This database was split into a training, a validation and a test set that respectively contained 12, 5 and 7 volumes.

B. U-net optimization set-up

The segmentation problem was solved by training a U-net CNN [7] which included batch normalization [11] and ReLU activation.

The network weights were initialized according to a uniform distribution with Xavier initializer [9] and the bias were initialized to 0.01.

The network parameters were optimized performing stochastic batch gradient descent with Adam optimizer [10]. Each batch was composed of 10 US sub-images of size 128×128. The gradient step was set to 1×10^{-4} during the first 10000 steps, 2×10^{-5} during the 10000 following steps and to 5×10^{-6} until the end of training.

Cross-entropy was defined as the loss for the first 5000 steps and softDice, as defined by Equation.1, was then used until the end of training.

$$\mathcal{L}(Y,\hat{Y}) \quad = \quad 1 - 2 \times \frac{\sum_{i=1}^{N} y_i \hat{y}_i}{\delta + \sum y_i + \hat{y}_i} \tag{1}$$

$y_i \in \{0,1\}$ is the label of the i-th pixel of the reference segmentation Y, 1 corresponding to the CVS. $\hat{y}_i \in [0,1]$ is the probability, given by the network output $\hat{Y}$, of the i-th pixel of X to belong to the CVS. δ was set to 10^{-10} to avoid division by zero when the batch did not contain CVS.

If the loss calculated over the validation set had not improved for 5000 iterations and if the network was trained for at least 15000 iterations, the training process was stopped (early stopping).

The network was implemented on python using tensorflow [13].

C. Training-set size

The aims of this experiment were to estimate the training-set size which is necessary to obtain a good precision ($Dice > 0.8$) and to determine which strategy should be used to keep improving the network accuracy increasing the training-set size or designing a more suitable architecture than U-net.

To address these questions, we fixed the validation set and the test set as defined in Section.II-A and trained a U-net with different training-set size : 1, 3, 5, 7, 9, 11 and 13 volumes. For each of these sizes, the training volumes were randomly drawn inside the training set and the experiment was repeated 10 times.

D. Artificial data augmentation

Increasing the training-set size in order to improve segmentation accuracy can be very difficult and time-consuming. Rather than manually annotating new data, artificial data augmentation strategies can be used to increase the variability contained in the training set. Nevertheless, the kind of transformation that should be performed is application dependent and must be chosen carefully. For this experiment, we tried :

1) Horizontal flip (betweeen the left and the right brain hemispheres).
2) Rotation (-10^o to 10^o).
3) Affine transformations of the pixel coordinates by a matrix A :
$$A = \begin{pmatrix} 1+m_{00} & m_{01} \\ m_{10} & 1+m_{11} \end{pmatrix} \tag{2}$$
with m_{00}, m_{01}, m_{10}, $m_{11} \in [-0.15, 0.15]$.
4) Non-linear transformations of the pixel coordinates by a gaussian random field.
5) Gaussian white noise ($\sigma^2 = 0.0625$).
6) Zoom (factor 1 to 3).

Each transformation was performed separately and was applied to each patch with probability 0.2. For each type of data augmentation, U-net was optimized 10 times. Examples of data before and after transformation can respectively be seen in Figure.1.a and Figure.1.b.

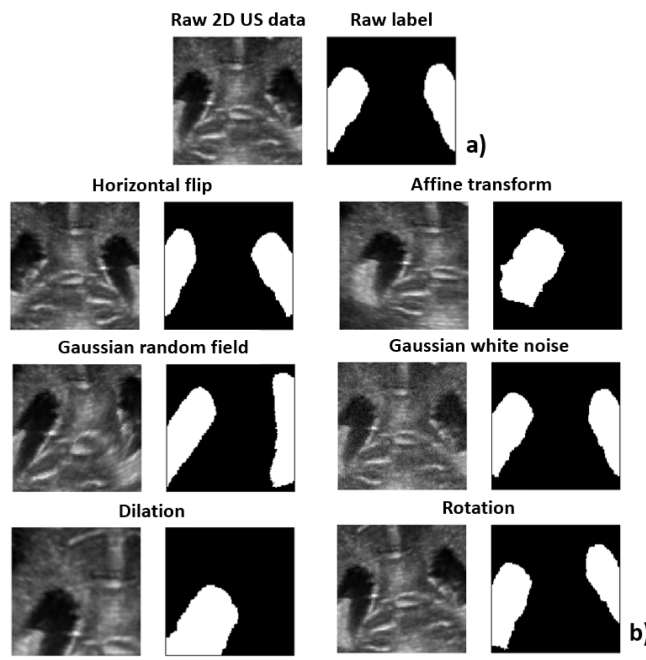

Fig. 1. US data and associated label without (a) and with artificial data augmentation (b).

E. Evaluation metrics

The precision of the trained networks were estimated over the test set with Dice as defined by Equation.3.

$$Dice \quad = \quad 2 \times \frac{|Y \cap \hat{\hat{Y}}|}{Y + \hat{\hat{Y}}} \tag{3}$$

Where Y is the targeted 3D segmentation, $\hat{\hat{Y}}$ the segmentation estimated by the trained network and $|X|$ a function that gives the number of 1 in a set X.

III. Results

A. Training set-size

The mean Dice (over 10 optimizations) obtained over the training set and the test set after the optimization of a U-net with different training-set size are represented in Figure.2.

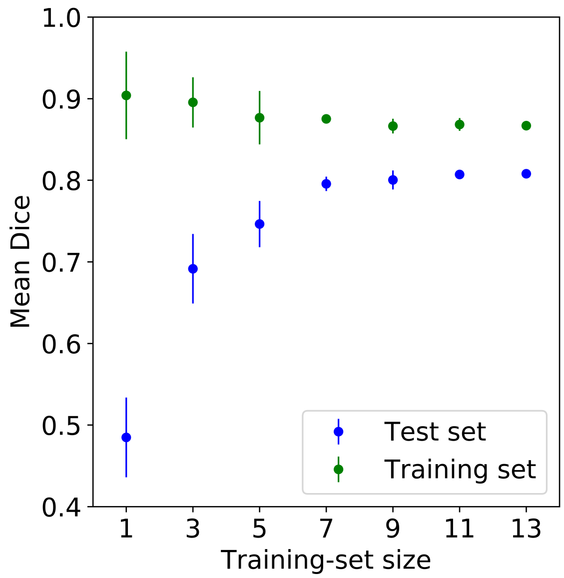

Fig. 2. Mean dice calculated over the training set and the test set as a function of the training-set size (without data augmentation).

We can see that the mean Dice calculated over the test set and the training set are respectively increasing and decreasing as a function of the training-set size. This behavior is coherent because the variability contained in the training-set increases with its size. Thus, the training set is harder to fit but enables a better generalization to unseen data.

When the training-set size is greater than 7, we can see little variation in the mean Dice values. To reveal a significant difference between the segmentation results obtained with different training-set size, t-test were performed. The results obtained with 13 training volumes were used as a baseline. Quantitative results and p-values are given in Table.I.

TABLE I
MEAN DICE VALUES OBTAINED AFTER THE OPTIMIZATION OF A U-NET WITH DIFFERENT TRAINING-SET SIZE.

Training patients	Mean Dice (training)	Mean Dice (test)	p-value
1	0.904 ± 0.054	0.485 ± 0.049	1.22e-13
3	0.895 ± 0.031	0.692 ± 0.043	1.97e-07
5	0.877 ± 0.033	0.746 ± 0.028	4.81e-06
7	0.875 ± 0.006	0.796 ± 0.009	2.08e-03
9	0.866 ± 0.009	0.8 ± 0.012	**9.10e-02**
11	0.868 ± 0.008	0.807 ± 0.004	**6.85e-01**
13	0.867 ± 0.006	0.808 ± 0.005	1

A good accuracy was obtained over the test set with only 9 training volumes (and 5 validation volumes). These results suggest that even if the creation of such database is time-consuming it remains worth considering for this problem.

A significant difference was found with the baseline for a training-set size of 1, 3, 5 and 7 volumes but not for a training-set size of 9 volumes ($p = 0.091 > 0.05$) or 11 volumes ($p = 0.685 > 0.05$). These results suggest that continuing to increase the training-set size would faintly improve U-net accuracy. It would require many new volumes or uncommon volumes to increase the variability contained in the training set and thus improve the segmentation accuracy. A better way to improve the network precision would be to design a CNN architecture that is more suitable for this problem.

B. Artificial data augmentation

The mean Dice (over 10 optimizations) obtained using each type of artificial data augmentation are given in Table.II. t-test were performed to reveal significant differences between the results obtained with and without artificial data augmentation.

TABLE II
SEGMENTATION ACCURACY OF A U-NET OPTIMIZED WITH DIFFERENT TYPES OF ARTIFICIAL DATA AUGMENTATION.

Augmentation	Dice	p-value
No augmentation	0.809 ± 0.003	1
Horizontal flip	**0.817 ± 0.003**	**3.42e-05**
Affine transform	0.811 ± 0.006	5.06e-01
Gaussian random field	0.811 ± 0.005	3.52e-01
Gaussian white noise	0.806 ± 0.003	1.12e-01
Rotation	0.811 ± 0.005	3.62e-01
Dilation	0.805 ± 0.009	2.06e-01

The use of horizontal flip during the optimization process significantly ($p = 8.84 \times 10^{-4}$) improved the segmentation accuracy in light of the Dice metric. This result might be explained by the fact that, as the left and right hemispheres are symmetric, performing horizontal flips enables the creation of realistic new data. Unfortunately, the other transformations did not significantly improve U-net accuracy. This result can be explained by the fact that these transformations did not enable the creation of realistic new US images : damage speckle, unrealistic anatomical orientation, unrealistic surface reverberation, . . . Nevertheless, the Mean Dice was higher with the use of affine transform, Gaussian random field and rotation.

C. 3D segmentation of the CVS

Qualitative results in the case of the patients which correspond to the highest Dice among the dilated CVS cases and the lowest and highest Dice among the normal CVS cases are respectively shown in Figure.3.a, Figure.3.b and Figure.3.c. These results were obtained with the U-net trained without data augmentation and the U-net trained with horizontal flips, which best performed on the validation set.

The main parts of the CVS were found in the three cases. No significant differences can be observed between the segmentation obtained with and without data augmentation in the case of the patient with the highest Dice among the dilated cases. Which ones are usually well segmented because all the parts of the ventricles are dilated and thus are easier to identify. In the normal cases, when U-net is trained with horizontal flips,

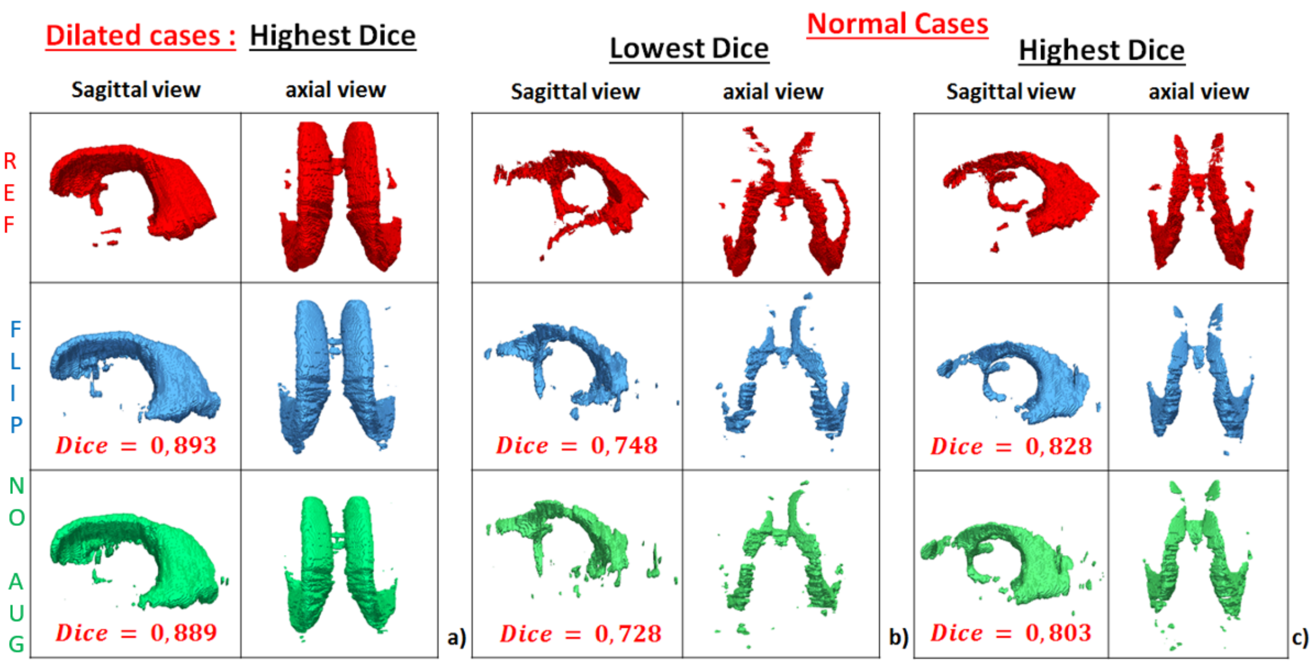

Fig. 3. Reference segmentation (red) and automatic segmentation obtained with (blue) and without (green) performing horizontal flip during the optimization process. The segmentation results correspond to the highest Dice among the dilated CVS (a) and the lowest (b) and highest (c) Dice among the normal CVS.

we can see that the number of outliers is reduced and that the temporal horns are better found.

IV. Conclusion

In this paper, a 2D U-net was trained to achieve CVS segmentation in 3D US data. This study aimed at estimating the number of training volumes that are necessary to obtain accurate segmentation results ($Dice > 0.8$) and at determining which types of artificial data augmentation are suitable for this application.

The experiments detailed in this paper showed that accurate segmentation can be obtained with only 9 training volumes and 5 validation volumes. In addition, we found that performing horizontal flips on the training data during the optimization process can significantly improves the network accuracy.

Acknowledgment

This work was performed within the framework of the LABEX CELYA (ANR-10-LABX-0060) and PRIMES (ANR-11-LABX-0063) of Universite de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). We would like to thank Dr. Borhane Slama, the innovation and clinical research commission president of CH Avignon, and the CH Avignon direction for their support to this research.

References

[1] Maunu, J., et al.: Ventricular dilatation in relation to outcome at 2 years of age in very preterm infants: a prospective Finnish cohort study. Developmental Medicine & Child Neurology, 53(1), 48-54 (2011).

[2] Qiu, Wu, et al.: Automatic segmentation approach to extracting neonatal cerebral ventricles from 3D ultrasound images. Medical image analysis 35 : 181-191 (2017).

[3] Boucher, M. A., Lippé, S., Dupont, C., Knoth, I. S., Lopez, G., Shams, R., ... & Kadoury, S. (2018). Computer-aided lateral ventricular and brain volume measurements in 3D ultrasound for assessing growth trajectories in newborns and neonates. Physics in Medicine & Biology, 63(22), 225012.

[4] Martin, M., Sciolla, B., Sdika, M., Wang, X., Quetin, P., & Delachartre, P. (2018, October). Automatic segmentation of the cerebral ventricle in neonates using deep learning with 3d reconstructed freehand ultrasound imaging. In 2018 IEEE International Ultrasonics Symposium (IUS) (pp. 1-4). IEEE.

[5] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

[6] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[7] Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pp. 234-241 (2015).

[8] Milletari, F., et al.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Fourth International Conference on 3D Vision, pp. 565-571 (2016).

[9] Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256).

[10] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[11] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

[12] Solberg, O. V., Lindseth, F., Torp, H., Blake, R. E., & Hernes, T. A. N. Freehand 3D ultrasound reconstruction algorithms—a review. Ultrasound in medicine & biology, 33(7), 991-1009 (2007).

[13] Martín Abadi & al. TensorFlow: Large-scale machine learning on heterogeneous systems (2015). Software available from tensorflow.org.