Deep Learning applied to multi-structure segmentation in 2D echocardiography : a preliminary investigation of the required database size

Sarah Leclerc
CREATIS
Universite de Lyon
Lyon, France

Erik Smistad
CIUS
NTNU
Trondheim, Norway

Thomas Grenier
CREATIS
Universite de Lyon
Lyon, France

Carole Lartizien
CREATIS
Universite de Lyon
Lyon, France

Andreas Ostvik
CIUS
NTNU
Trondheim, Norway

Florian Espinosa
Cardiovascular department
University-affiliated hospital of St-Etienne
Saint-Etienne, France

Pierre-Marc Jodoin
Computer Science department
University of Sherbrooke
Sherbrooke, Canada

Lasse Lovstakken
CIUS
NTNU
Trondheim, Norway

Olivier Bernard
CREATIS
Universite de Lyon
Lyon, France

Abstract—**With the recent advances in machine learning, and their successful application to medical imaging, building medical databases big enough to learn to solve corresponding tasks for any given patient has become a priority. In this study, we set up a specific dataset of 500 patients to investigate the number of patients needed by two learning methods, the Structured Random Forest (SRF) and U-net, to obtain accurate segmentation results in 2D echocardiography. Our findings advocate that U-net is a good candidate to solve the still-open 2D echocardiography automatic segmentation problem.**

Index Terms—**2D Echocardiography, Structured Random Forests, Machine learning, Multiclass segmentation**

I. Introduction

Fast, cheap and safe to use, echocardiography is to this day the primary clinical imaging modality to analyze the cardiac function. Cardiologists establish their diagnosis based on several clinical indices, including the Ejection Fraction (EF) of the left ventricle (LV) and the Global Longitudinal Strain. Accurate segmentation of the left ventricle border - the endocardium (endo) - and of the myocardium external border - the epicardium (epi) - enable to estimate EF and GLS from 2D acquisitions.

A. Motivation

In daily practice, cardiologists segment 2D acquisitions of apical 2 chambers (2CH) and 4 chambers (4CH) views at end-diastole (ED) and end-systole (ES) with semi-supervised algorithms. These methods require manual interactions to perform a precise delineation of the heart, which makes the process time consuming, non-consensual and hard to reproduce. The complete and robust automation of cardiac segmentation has therefore been a dynamic research field.

Segmentation in echocardiography is especially difficult due to the quality of ultrasound images, characterized by a low contrast, speckled textures, acoustic shadows and reverberation artefacts. Learning from expertise appears necessary to cope with the inherent variability of the task. Convolutional Neural Networks (CNN) have shown their potential in reproducing experts actions on several medical applications [1]. However, the amount of training data needed for a CNNs to achieve competitive results is unclear.

This quantity being task dependant, we here focus our inquiry on 2D ultrasound segmentation of the heart. To this end, we built a suitable dataset from daily clinical acquisitions to investigate the capacity of a U-net neural network to reproduce expert contouring of the endocardium and epicardium.

B. Related work

Cardiac segmentation has in the past been approached with bottom-up algorithms, active contours and atlas registration algorithms [2] [3] [4]. These methods often required the supervision of an expert.

Data-based driven techniques have become more and more popular in the medical research community [5], though their use was constrained by the unavailability of clinical data. In cardiac segmentation, MICCAI challenges such as the Challenge on Endocardial Three-dimensional Ultrasound Segmentation (CETUS) in 2014 [5] and the Automated Cardiac Diagnosis Challenge

(ACDC) in 2017 [6] provided public datasets in 3D echocardiography and MRI. Deep learning segmentation methods, and U-net architectures in particular, proved to be promising solutions for training sets of respectively 200 and 30 images [8] [7].

As for 2D echocardiography, it was confirmed in a previous study [9] that the fully automatic Machine Learning technic called Structured Random Forests could provide similar results as state-of-the-art semi-automatic Active Appearance Models on a dataset of 200 4CH views at ED and ES. To train SRF, the features used to describe the image space have to be manually picked, which we identified as their main limitation. The impact of increasing the training set size of a U-net was investigated in [10], where it involved generated segmentations of another algorithm and not experts contouring.

C. Contributions

This study provides two main contributions :

- Give insights as to how a CNN responds to an increasing training set size.
- Compare its performance to SRF, another fully automatic learning method.

This experiment is part of a larger analysis of supervised learning methods addressing the design of a robust automatic multi-structure segmentation solution.

II. Methods

We compare two models, a U-net and a Structured Random Forest, both optimized by tuning hyperparameters by cross-validation on the whole dataset.

A. U-net

U-net is a fully convolutional neural network introduced in [12] that can be trained to reconstruct a segmentation map from a low-dimensional representation of the input. Skip-connections between the encoding and decoding part enable to use features at several resolutions. Its simplicity and effectiveness have made it popular inside and outside the medical community.

As shown in Fig 1, our version of U-net is a bit different from the original one as it incorporates :

- Batch Normalization
- Spatial Dropout at 20%
- concatenation as skip-connections (no cropping of feature maps)
- no data augmentation

ReLU functions used as activations after all Batch Normalization layers, except the last where we apply a softmax to get a distribution across classes. Each pixel is given the class with the highest coefficient. A single model is trained to segment echocardiographic 2D images regardless of view and instant.

Our U-Net architecture

Level	Layer	Kernel/Pooling size	Skip connection
D1	Conv	32*(3,3)	1
	BatchNorm		
	Conv	32*(3,3)	
	BatchNorm		
	MaxPooling	(2*2)	
D2	Conv	64*(3,3)	2
	BatchNorm		
	Conv	64*(3,3)	
	BatchNorm		
	MaxPooling	(2*2)	
D3	Conv	128*(3,3)	3
	BatchNorm		
	Conv	128*(3,3)	
	BatchNorm		
	MaxPooling	(2*2)	
D4	Conv	256*(3,3)	4
	BatchNorm		
	Conv	256*(3,3)	
	BatchNorm		
	MaxPooling	(2*2)	
D5	Conv	512*(3,3)	
	BatchNorm		
	Conv	512*(3,3)	
	BatchNorm		
U1	Deconv	256*(2,2) - stride(2,2)	4
	BatchNorm		
	Conv	256*(3,3)	
	BatchNorm		
	Conv	256*(3,3)	
	BatchNorm		
U2	Deconv	64*(2,2) - stride(2,2)	3
	BatchNorm		
	Conv	64*(3,3)	
	BatchNorm		
	Conv	64*(3,3)	
	BatchNorm		
U3	Deconv	64*(2,2) - stride(2,2)	2
	BatchNorm		
	Conv	96 (3,3)	
	BatchNorm		
	Conv	96 (3,3)	
	BatchNorm		
U4	Deconv	32*(2,2) - stride(2,2)	1
	BatchNorm		
	Conv	48 (3,3)	
	BatchNorm		
	Conv	48 (3,3)	
	BatchNorm		
Seg	Conv	4 (1,1)	

implemented with Keras [11]

Fig. 1. Number and size of filters are given for convolution layers, the stride for MaxPooling layers and ”Deconvolution” layers. The Skip Connection number pinpoints which layers are connected.

B. SRF

Structured Random Forests are a decision tree based method that aims at incorporating contextual information in pixel-wise classification. To do so, images are fragmented into patches and a segmentation patch is voted by several random trees. The segmentation map is reconstructed from overlapping patches, resulting in a smooth segmentation (see [9] for more details on SRF).

Main differences between SRF and U-net include:

- Selection of features : SRF features are handcrafted and not learnt end-to-end.
- Split function : SRF split the data into two subsets based on a single feature whereas U-net activations

rely on multiple features at each layer and does not perform routing of the data.

- Optimization : Trees are grown recursively based on a splitting gain criteria while U-net model are fixed and optimized with stochastic gradient descent.

Data is divided in subsets of 50 patients (200 images) and 12 random trees are built for each subset by relying on the Histogram of Gradients (HOG) computed at several resolutions for $2e5$ patches. We train a single SRF to segment ED / ES and 2CH / 4CH images.

III. Experiment

The dataset from [9] has been extended to 500 cases and is split into 10 folds of 50 patients each. We keep one for test, one for validation, and increasingly add folds into the training set of the U-net. The number of epochs is adapted so that models have the same number of weight updates. Other hyperparameters (learning rate, batch size etc...) are constant.

A. Metrics

We base our evaluation of the LV and myocardium segmentation on three geometric metrics :

- the Sorensen-Dice index, to assess the overlap of the regions S delimited by the endocardium and the epicardium : $D = 2\left|S_{ref} \cap S_{algo}\right| / (\left|S_{ref}\right| + \left|S_{algo}\right|)$
- the Mean Absolute Distance (MAD) between contours C to study the average performance : $MAD = \overline{\min(AD_{C_{ref}-C_{algo}}, AD_{C_{algo}-C_{ref}})}$, AD being the set of absolute distances obtained from projecting contour points onto the second contour.
- the Hausdorff distance between contours to study the extreme error : $HD = \max(\min(AD_{C_{ref}-C_{algo}}, AD_{C_{algo}-C_{ref}}))$

B. Results

Statistical results are illustrated with Tukey boxplots. They summarize distributions by showing the median, lower and upper quartiles. The whiskers are set at $1.5 * IQR$, with IQR the inter-quartile range.

1) U-net: We plot the boxplots of the eight models learnt on a growing training dataset. Each boxplot corresponds to the performance on the test set.

2) Comparison with SRF: We provide the MAD boxplot of SRF as well as tables comparing the two algorithms for the smallest and biggest training dataset.

IV. Observations

A. Mean performances

As seen in Fig. 2, 3 and 4, the U-net model shows a distinct improvement when increasing the training size on the three metrics and on both structures. This improvement is not linear nor consistent. For example :

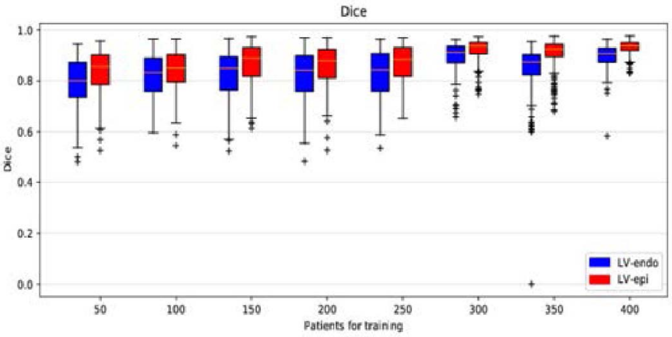

Fig. 2. D evolution when increasing the training set size of our U-Net.

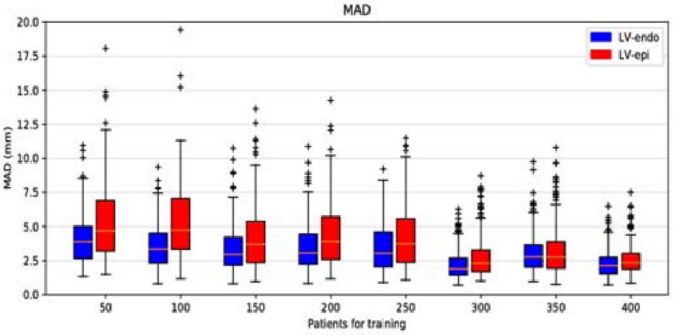

Fig. 3. MAD evolution when increasing the training set size of U-Net.

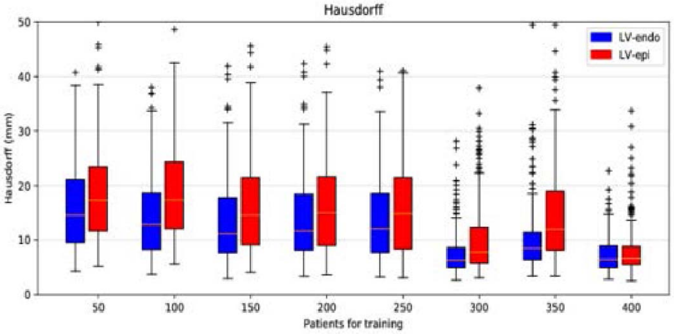

Fig. 4. HD evolution when increasing the training set size of U-Net.

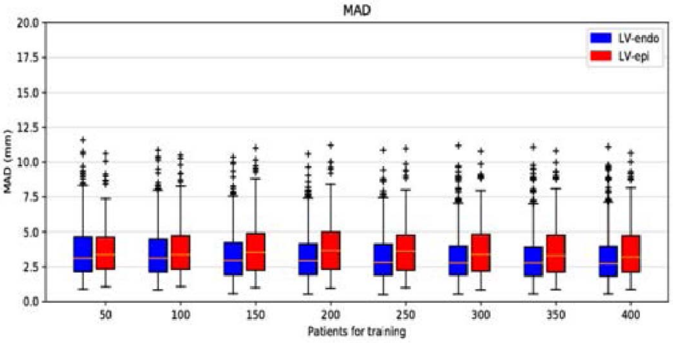

Fig. 5. MAD evolution when increasing the training set size of SRF.

TABLE I
COMPARISON BETWEEN U-NET AND SRF
TRAINING SIZE = 50 PATIENTS / 200 IMAGES

Model	LV-endo			LV-epi		
	D	MAD	HD	*D*	MAD	HD
U-net	0.795 ±**0.089**	4.1 ±**1.9**	15.7 ±**7.6**	0.838 ±0.083	5.3 ±2.8	18.3 ±**8.5**
SRF	**0.843** ±0.095	**3.7** ±2.1	**13.9** ±9.5	**0.894** ±**0.058**	**3.7** ±**1.9**	**14.6** ±9.0

TABLE II
COMPARISON BETWEEN U-NET AND SRF
TRAINING SIZE = 400 PATIENTS / 1600 IMAGES

Model	LV-endo			LV-epi		
	D	MAD	HD	*D*	MAD	HD
U-net	**0.896** ±**0.047**	**2.3** ±**1.0**	**7.3** ±**3.2**	**0.931** ±**0.028**	**2.6** ±**1.1**	**8.1** ±**4.7**
SRF	0.859 ±0.092	3.3 ±2.1	12.7 ±9.6	0.896 ±0.06	3.7 ±2.0	14.2 ±9.0

- Going from 250 to 300 leads to a much higher gain than any other subset.
- Going from 300 to 350, there is actually a decrease in performance.

With only 50 cases, the SRF show in table I systematically better mean performances, however as can be observed in Fig. 5, adding training cases is not as beneficial for this method as for the U-net. With 400 training cases in table II, the U-net gets higher scores with a significant margin.

B. Robustness

Overall, the boxplots 2, 3 and 4 suggest that the robustness of our U-Net gets much better as we add training cases, as can be seen from the U-net standard deviation values in tables I and II. Outliers are also reduced in number and in acuteness.

The SRF method on the other hand, and as was concluded in [9], displays good mean performances even on a small dataset (see table I), but they remain unconsistent as suggested by the standard deviation values (Fig.5).

V. DISCUSSION

The U-net behavior suggests that beyond data quantity, data diversity is necessary to ensure generalization and that the performance assessment is biased by the validation and test sets used to select and evaluate the model. 200 patients were sufficient to outperform SRF as regards to the MAD on both structures. The U-net is able to further improve by adding training data but seems to slowly reach convergence.

The SRF appear to reach a plateau that is not linked to the training dataset size. It is our belief that using handcrafted features simplifies the problem by restricting access to only relevant information, but also limitates the information that could be extracted from the raw ultrasound image. Following this idea and as we observed, such algorithm would perform well even with a small training dataset, but wouldn't benefit from more data as neural networks would.

VI. CONCLUSION

U-net models are able to integrate the variability of additional training data. With only 50 cases, our CNN gave competitive results on cardiac ultrasound segmentation, and from 300 patients on, it outperformed the SRF algorithm. As the number of outliers and standard deviation values are still high, further inquiries will focus on improving its robustness by adding shape constraints to encourage predictions of anatomically plausible shapes.

ACKNOWLEDGMENT

This work was supported by the LABEX PRIMES program (ANR-11-LABX-0063) of Universite de Lyon, operated by the French National Research Agency (ANR).

REFERENCES

[1] Geert J. S. Litjens, Thijs Kooi, Babak Ehteshami Bejnordi et al., "A Survey on Deep Learning in Medical Image Analysis", Medical image analysis, vol 42, pp.60-88, 2017.

[2] J. G. Bosch, S. C. Mitchell, B. P. F. Lelieveldt et al., "Automatic segmentation of echocardiographic sequences by active appearance motion models," IEEE Trans. Med. Imag., vol. 21, no. 11, pp. 1374–1383, Nov. 2002.

[3] N. Lin, W. C. Yu, J. S. Duncan, "Combinative multi-scale level set framework for echocardiographic image segmentation," Med. Image Anal.,' vol. 7, no. 4, pp. 529-537, Dec. 2003.

[4] O. Oktay, A. Gomez, K. Keraudren, A. Schuh, W. Bai et al, "Probabilistic Edge Map (PEM) for 3D Ultrasound Image Registration and Multi-atlas Left Ventricle Segmentation," Functional Imaging and Modeling of the Heart. Maastricht, The Netherlands: Springer; pp. 223230; 2015.

[5] O. Bernard, J.G. Bosch, B. Heyde et al., "Standardized Evaluation System for Left Ventricular Segmentation Algorithms in 3D Echocardiography," in IEEE Transactions on Medical Imaging, vol. 35, no. 4, pp. 967-977, April 2016.

[6] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky and others, "Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved?," IEEE Transactions on Medical Imaging, 2018

[7] O. Oktay, E. Ferrante, K. Kamnitsas et al., "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation", IEEE Transactions on Medical Imaging, vol 37, no. 2, pp 384-395, Sept 2017.

[8] C.Zotti, Z.luo, A.Lalande et P.Jodoin, "Convolutional Neural Network with Shape Prior Applied to Cardiac MRI Segmentation", IEEE Journal of Biomedical and Health Analysis, 2018

[9] S. Leclerc, T. Grenier, F. Espinosa, O. Bernard, "A fully automatic and multi-structural segmentation of the left ventricle and the myocardium on highly heterogeneous 2D echocardiographic data," IEEE International Ultrasonics Symposium (IUS), 2017

[10] E. Smistad, A. Ostvik, B. O. Haugen, L. Lovstakken, "2D left ventricle segmentation using deep learning", IEEE International Ultrasonic Symposium (IUS), 2017

[11] Chollet, François and others, "Keras", 2015, https://keras.io

[12] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional networks for biomedical image segmentation", Proc. Med. Image Comput. Comput.-Assist. Intervention, pp. 234-241, 2015.