AUTOMATIC CAROTID ULTRASOUND SEGMENTATION USING DEEP CONVOLUTIONAL NEURAL NETWORKS AND PHASE CONGRUENCY MAPS

Carl Azzopardi, Yulia A. Hicks

Cardiff School of Engineering
Cardiff University, Wales

Kenneth P. Camilleri

Centre for Biomedical Cybernetics
University of Malta, Malta

ABSTRACT

The segmentation of media-adventitia and lumen-intima boundaries of the Carotid Artery forms an essential part in assessing plaque morphology in Ultrasound Imaging. Manual methods are tedious and prone to variability and thus, developing automated segmentation algorithms is preferable. In this paper, we propose to use deep convolutional networks for automated segmentation of the media-adventitia boundary in transverse and longitudinal sections of carotid ultrasound images. Deep networks have recently been employed with good success on image segmentation tasks, and we thus propose their application on ultrasound data, using an encoder-decoder convolutional structure which allows the network to be trained end-to-end for pixel-wise classification. Concurrently, we evaluate the performance for various configurations, depths and filter sizes within the network. In addition, we further propose a novel fusion of envelope and phase congruency data as an input to the network, as the latter provides an intensity-invariant data source to the network. We show that this data fusion and the proposed network structure yields higher segmentation performance than the state-of-the-art techniques.

Index Terms— Ultrasound, Segmentation, Deep Convolutional Networks, Carotid Artery, Phase

1. INTRODUCTION

Cerebrovascular disease is amongst the leading causes of death in the United States [1]. The underlying cause is atherosclerosis - a vascular pathology which is characterised by the thickening and hardening of blood vessel walls [2]. The carotid is one such artery which is susceptible to atherosclerotic deposits - or plaque. When plaque ruptures in the carotid artery, there is a significant risk that the blood clot which forms will travel upstream to occlude a narrower vessel in the brain - ultimately leading to a stroke [3, 4].

Localisation and grading of the severity of a stenosis forms a large part of the diagnostic process used to assess the risk of rupture. While techniques such as Digital Subtraction Angiography and Magnetic Resonance Angiography are presently considered to be the gold standard in assessing carotid disease [5, 6, 7], Ultrasound Imaging has increasingly gained popularity due to its low cost and non-invasive nature, permitting a quick assessment of vessel geometry, degree of stenosis and plaque morphology [3, 5, 7, 8]. In order for accurate assessment of plaque morphology and burden to take place using metrics such as Total Plaque Volume (TPV) or Vessel Wall Volume (VWV), two specific wall interfaces need to be identified: the media-adventitia boundary (MAB) and the Lumen - Intima boundary (LIB) [9]. Delineation needs to be robust and reproducible, and manual methods have been shown to be tedious, labour intensive [9], and prone to variability [10]. Automated or semi-automated segmentation algorithms which facilitate this process are therefore required.

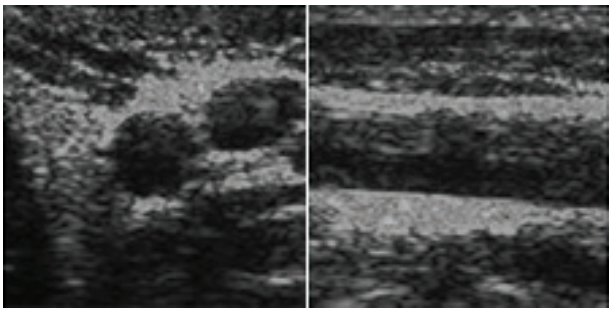

Fig. 1. An example of a transverse [left] and longitudinal [right] carotid ultrasound image.

A number of studies have looked into the segmentation of wall interfaces in transverse carotid ultrasound images [9, 11, 12], although these have not fully addressed the issue of manual tuning of technique parameters to achieve a good segmentation. In this work, we propose the use of Deep Convolutional Neural Networks (DCNNs) for delineating the MAB in healthy subjects from fused envelope and phase congruency data. Future research will also evaluate the performance of using a two-stage DCNN in delineating both the MAB and LIB in symptomatic subjects displaying significant carotid stenosis. Segmentation of structures in images may be interpreted as a pixel classification problem, and DCNNs

978-1-5090-1172-8/17/$31.00

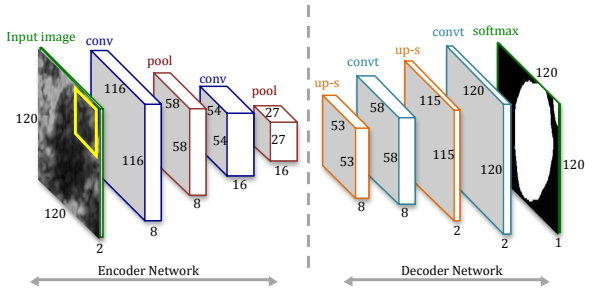

Fig. 2. A graphical representation of the encoder-decoder structure of the Deep Neural Network.

can provide such a classification by constructing hierarchies of features from training data [13]. Previous work exists that applies DCNNs for medical image segmentation, although these focus on different applications [13, 14], or else use different structures to classify pixels within sub-image patches to estimate the I-M Thickness in longitudinal images [15].

Our approach proposes to use DCNNs with an encoder-decoder structure to predict segmentation masks from image data [16]. This is a faster approach, applying end-to-end pixel classification on entire images, as opposed to typical patch-based methods. We explore various network structures, depths and filter sizes, to evaluate performance on both transverse and longitudinal images. We further propose the use of phase congruency maps as an input to the DCNN, as these provide a feature space which is robust to variations in signal intensity [17], and we finally show that a fusion of amplitude data and this phase congruency data as input to the DCNN yields an improved segmentation performance.

2. METHODOLOGY

2.1. Data

A total of 250 transverse and 250 longitudinal images were obtained from 5 subjects with normal (non-stenotic) carotids, with ages spanning 25 - 40 years. Images were acquired using an Ultrasonix Sonix RP Ultrasound machine (Analogic Corporation, Peabody, MA, USA), equipped with a 14 MHz L14-5 Linear Probe. Images were not normalised and manually traced independently by two radiographers. Each radiographer traced the image sets twice, with a period of 2 weeks in between sessions. The MAB was traced in both transverse and longitudinal sections. From all the image set available, a random selection of 70% were kept as training data, 20% as validation data, and 10% as testing data. Images were manually cropped to 120×120 px, centred on the artery, thus mimicking a typical machine's GUI windowing function. Dataset size was augmented by rotating images and their corresponding labels through 90, 180 and 270 degrees. The rotated data was then appended to the original data, creating a total of 1000 longitudinal and 1000 transverse images, as well as labelled data sets for each.

2.2. Extracting Phase Congruency Maps

Speckle noise, low contrast and local changes in intensity make ultrasound image segmentation an inherently difficult problem for methods seeking to delineate contours of interest. Different end-user preferences on intensity gain settings further exacerbates the problem, making it difficult to find optimal parameters in a segmentation method which apply across the board. Phase information provides a feature space which is theoretically invariant to amplitude changes, and preserves structural information of a signal. The local energy model, proposed by [18], postulates that when Fourier components of an image are maximally in phase, and thus *phase congruency (PC)* reaches a maximum, features may be perceived, thus yielding an interesting alternative data representation. The authors in [18] define phase congruency as:

$$PC(x) = \max_{\bar{\phi}(x)\in[0,2\pi]} \frac{\sum_n A_n \cos(\phi_n(x) - \bar{\phi})}{\sum_n A_n} \quad (1)$$

where, if we consider the Fourier series expansion of a 1D signal, $I(x) = \sum_n A_n cos(\phi_n(x))$, then A_n is the amplitude of the *n*th Fourier component, and $cos(\phi_n(x))$ is the *n*th co-sinusoidal harmonic having some phase ϕ. The value of $\bar{\phi}(x)$, over which the equation is maximised, is the amplitude weighted mean local phase angle of all Fourier components at the point being considered. In our case, we compute the 2D PC maps for each image using a more convenient method by Kovesi in [19], whereby the PC maps are obtained from the even and odd responses of a filter bank of quadrature logarithmic Gabor filters convolved with the signal. We then compute the maximum moment of the PC maps as an indication of feature significance, and use this as a secondary channel of information to the regular B-mode amplitude data, when submitting this to the DCNN.

2.3. Deep Convolutional Neural Networks

A DCNN is a multilayer perceptron network which can exploit the stationary nature of natural images by learning features on locally connected pixels. The convolutional layers learn small features from small image patches sampled from the whole image [13]. The sub-sampling layers are used to reduce the computational complexity by summarising the statistics of a feature over a region in the image [13]. Our image segmentation task may be posed as a pixel-by-pixel classification problem, whereby a decision is made for each pixel - classifying it into 'foreground' or 'background'. The output of the network will therefore be a segmentation mask, ideally matching the manual segmentation (ground truth) provided by the expert. This may be defined as an optimisation problem, whereby we attempt to minimize the error between our

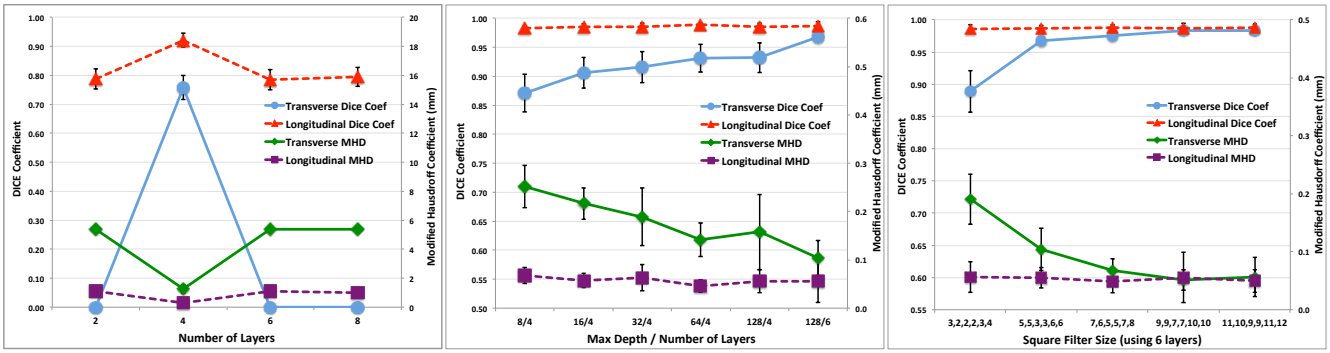

Fig. 3. DICE Coefficient and MHD performance for no. of layers [left]; depth of output maps [middle] and filter size [right].

output mask and the ground truth, by finding the optimal set of parameters for the objective function:

$$J(\theta, b) = \sum_{k=1}^{m} (h(x^{(k)}; \theta, b) - y^{(k)})^2 \tag{2}$$

The objective function $J(\theta, b)$ is the sum of differences between the predicted output segmentation mask $h(x)$ and the ground truth labels $y(x)$, over all different training examples $k = 1, 2, 3...m$. The mask $h(x)$ is the result of a feed forward operation carried out through the network. Referring to Figure 2, we note that the full DCNN is constructed using two parts: an encoder section and a decoder section. The encoder is built using stacks of convolutional and subsampling layers, whereas the decoder is built using stacks of convolutional transpose layers and up-sampling layers. Typical CNNs normally resemble just the encoder structure, with their final layer being fully interconnected to a one-dimensional layer of nodes, before feeding on to the output. Such CNNs however have the inherent drawback of loss of image resolution, arising partially out of the convolution operation in the convolution layers, and partially out of the subsampling process designed to summarise the feature space. Since we intend to train our network in an end-to-end fashion, a decoder is appended to the end of the encoding CNN, to expand the extracted features back to full resolution, while concurrently calculating a probabilistic mask. The feedforward equation for a particular layer l in the encoder is provided by the following deterministic function $g^{(l)}$:

$$\begin{aligned} z_j^{(l)} &= g^{(l)}(z_i^{(l-1)}, w_{ij}^{(l)}, b_j^{(l)}) \\ &= \psi\left(\rho\left(\sum_{i=0}^{I} \tilde{w}_{i,j}^{(l)} * \ z_i^{(l-1)} + b_j^{(l)}\right)\right) \end{aligned} \tag{3}$$

where $z_j^{(l)}, j \in [1, F]$ is the j^{th} output feature map for layer l, calculated by convolving the trainable convolution filter $w_{ij}^{(l)}$ with the input to that layer $z_i^{(l-1)}$. The index i denotes the number of input maps available from the preceding layer, F denotes the number of filters, $b_j^{(l)}$ denotes the trainable bias term for layer l, $\tilde{w}$ denotes the flipped version of w [16], and $*$ denotes the convolution operator. The function $\rho(x)$ denotes the rectified linear activation function (ReLu), defined as $\rho(x) = max(0, x)$, whereas the function $\psi(x)$ is used to define the sub-sampling function. Subsampling functions normally implement either a *max pooling* function, whereby the maximum value from the preceding layer of local connections is passed onwards, or a mean pooling function, whereby the average is passed onwards to the next layer instead of the maximum. Within the decoder structure, the feedforward equation is provided by the function $h^{(l)}$ [20]:

$$\begin{aligned} y_j^{(l)} &= h^{(l)}(y_i^{(l-1)}, w_{ij}^{(l)}, b_j^{(l)}) \\ &= \rho\left(\sum_{i=0}^{I} w_{i,j}^{(l)} \circledast \Psi(y_i^{(l-1)}) + b_j^{(l)}\right) \end{aligned} \tag{4}$$

where $y_j^{(l-1)}$ in the first instance would be z from the preceding encoder layer. Thereafter it would be simply the output of the previous decoding layer. The function $\Psi(x)$ denotes an up-sampling operation, and the operator $\circledast$ refers to the transposed full convolution. Each layer is once again followed by a ReLu function $\rho(x)$. At the end of the decoder network, the number of output maps are reduced to two, and fed into a softmax classifier, which provides logistic regression for a two class problem [13]. The softmax function $\sigma(z)$ has the effect of maximising the maximum value of the outputs, making these close to 1, and the rest close to 0.

3. RESULTS AND DISCUSSION

Several experiments were carried out to test the effect of different structural DCNN hyperparameters on the overall segmentation performance. These were: a) the number of layer 'stacks', that is: the number of [convolution + subsampling] or [transpose-convolution + upsampling] stacks used between the input and final segmentation mask; b) the number of input / output maps utilized within each layer stack;

and c) the filter dimensions used within convolutional / convolutional transpose layers. Given that a grid search of all hyperparameters combinations yields a large and expensive parameter space to test, a heuristic approach was used. A single hyperparameter was varied, while keeping the others constant. Once the best performing hyperparameter value was identified, it was fixed, and the next hyperparameter varied. Cross-validation was implemented by keeping a hold-out test set that the training process never touches.

The DCNN was built using the MatConvNet toolbox and trained on an Intel Core i7 with a Geforce GT 650M video card. Training took 300 epochs using stochastic gradient descent, with batches of 5 ultrasound images, a learning rate α of 5e-7, a momentum of 0.90 and a weight decay of 5e-6. The weights and bias terms were randomly initialised. The DICE coefficient of similarity [9] and the Modified Hausdorff Distance (MDH) [21] were used as performance metrics, to compare to other studies which utilise different techniques for MAB segmentation. Figures 3a to 3c show the performance of the network using the DICE coefficient and MHD. Figure 3a shows that, while keeping a fixed square filter size of 5×5 for encoding and 6×6 for decoding, and single input-output maps at each stack, the ideal number of stacks is 4. We note however that once the input/output map depths at each layer are varied, the best performing number of stacks becomes 6, as shown in Figure 3b. For brevity, the x-axis here shows the greatest number of output maps utilised in the central stack, and the number of stacks utilised. Using 6 layer stacks with 128 output maps at the centre, the effect of varying the square filter size at each convolutional / deconvolutional layer is then explored. Figure 3c shows that performance increases with filter size, but that it then levels off at DICE / MHD values in the region of 0.988 and 0.050 mm respectively. These findings indicate that, in agreement with literature [22], increasing depth of stacks or output maps typically yields better results, as the network is able to account for non-linearities within data. For stack number, performance eventually levels off however, as the increased number of sub-sampling layers compromise the resolution of the inner layers.

The results shown in Figure 4a display the DICE coefficients obtained with the optimal network structure, while varying the source of input data. Results are shown for using amplitude data, phase congruency moment data, and fused amplitude-phase congruency data. We note that while amplitude data on its own provides better results than phase information, the latter yields reasonable DICE coefficients which are in excess of 0.95. Phase information has the advantage of being amplitude invariant, and thus theoretically provides consistent results regardless of user selected amplitude settings. The fusion of amplitude and phase in turn yields the best results across both transverse and longitudinal images. Comparisons against recent literature in [9, 12] and [23] are shown, whereby only the performance on transverse images were reported for the MAB segmentation performance. We note that DCNNs applied to amplitude information alone yields comparable performance to that obtained by Ukwatta *et al.* in [9], but that its combination with phase information yields superior results. Figure 4b and 4c show examples of the DCNN generated contours (green) in comparison to the average manually segmented contours (red).

4. CONCLUSION

The study has evaluated the use of a DCNN as a segmentation tool as well as the effects of its hyperparameters on segmentation performance. It has also demonstrated that the DCNN achieves performance which meets the state-of-the-art in MAB delineation when using amplitude data, and which exceeds state-of-the-art when amplitude information is combined with phase information. A recognised limitation of the study was the size of the dataset available to train the network. Future work will evaluate segmentation performance on a larger dataset and quantify plaque burden when using a modified DCNN to concurrently segment the MAB and LIB in patients displaying plaque and carotid stenosis.

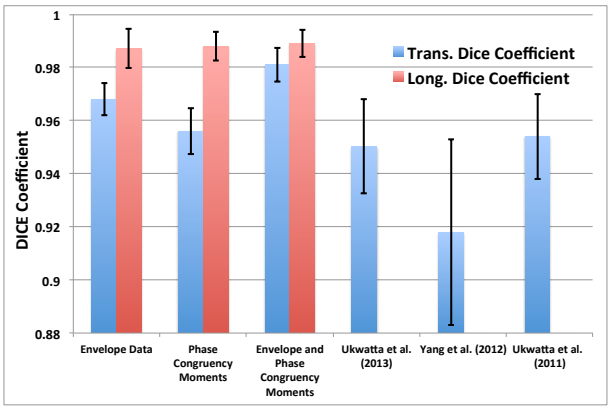

Fig. 4. Performance of envelope, phase congruency, and fused data on dice coefficients.

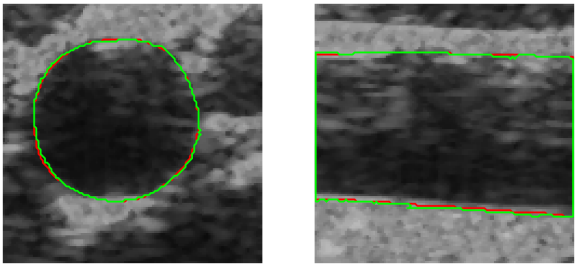

Fig. 5. Examples of segmented Transverse image [left] and longitudinal image [right]. Manual annotation is in red, while DCNN segmentation is in green.

5. REFERENCES

[1] Mozaffarian et al., *Heart disease and stroke statistics-2015 update : A report from the American Heart Association*, vol. 131, 2015.

[2] Ali Al-Mamari, "Atherosclerosis and physical activity.," *Oman medical journal*, vol. 24, no. 3, pp. 173–8, jul 2009.

[3] Zeynettin Akkus et al., "Fully automated carotid plaque segmentation in combined contrast-enhanced and B-mode ultrasound," *Ultrasound in Medicine and Biology*, vol. 41, no. 2, pp. 517–531, 2015.

[4] Jonathan Golledge, Roger M Greenhalgh, and Alun H Davies, "The symptomatic carotid plaque.," *Stroke; a journal of cerebral circulation*, vol. 31, pp. 774–781, 2000.

[5] A. Long, A. Lepoutre, Corbillon, and A. Branchereau, "Critical Review of Non- or Minimally Invasive Methods (Duplex Ultrasonography, MR- and CT-angiography) for Evaluating Stenosis of the Proximal Internal Carotid Artery," *European Journal of Vascular and Endovascular Surgery*, vol. 24, no. 1, pp. 43–52, jul 2002.

[6] Bruno Randoux et al., "Carotid artery stenosis: prospective comparison of CT, three-dimensional gadolinium-enhanced MR, and conventional angiography.," *Radiology*, vol. 220, no. 1, pp. 179–85, jul 2001.

[7] J D Gill, H M Ladak, D a Steinman, and a Fenster, "Accuracy and variability assessment of a semiautomatic technique for segmentation of the carotid arteries from three-dimensional ultrasound images.," *Medical physics*, vol. 27, no. 6, pp. 1333–42, 2000.

[8] André Miguel F Santos et al., "A novel automatic algorithm for the segmentation of the lumen of the carotid artery in ultrasound B-mode images," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6570–6579, 2013.

[9] E Ukwatta et al., "Three-dimensional ultrasound of carotid atherosclerosis: semiautomated segmentation using a level set-based method.," *Medical physics*, vol. 38, no. 5, pp. 2479–2493, 2011.

[10] Fei Mao, Jeremy Gill, Donal Downey, and Aaron Fenster, "Segmentation of carotid artery in ultrasound images: Method development and evaluation technique," *Medical Physics*, vol. 27, no. 8, pp. 1961, 2000.

[11] Jose C R Seabra et al., "A 3-D ultrasound-based framework to characterize the echo morphology of carotid plaques," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 5, pp. 1442–1453, 2009.

[12] Xin Yang et al., "Segmentation of the common carotid artery with active shape models from 3D ultrasound images," feb 2012, p. 83152H.

[13] Adhish Prasoon et al., "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network.," *MICCAI*, vol. 16, no. Pt 2, pp. 246–53, 2013.

[14] Hariharan Ravishankar et al., "Hybrid Approach for Automatic Segmentation of Fetal Abdomen from Ultrasound Images using Deep Learning," in *ISBI*, 2016, pp. 779–782.

[15] R M Menchon-Lara and J L Sancho-Gomez, "Fully automatic segmentation of ultrasound common carotid artery images based on machine learning," *Neurocomputing*, vol. 151, pp. 161–167, 2015.

[16] Tom Brosch et al., "Deep Convolutional Encoder Networks for Multiple Sclerosis Lesion Segmentation," in *Lecture Notes in Computer Science*, Nassir Navab et al., Eds., vol. 9351 of *Lecture Notes in Computer Science*, pp. 3–11. Springer International Publishing, Cham, 2015.

[17] M Mulet-Parada and J A Noble, "2D+T acoustic boundary detection in echocardiography.," *Medical image analysis*, vol. 4, no. 1, pp. 21–30, 2000.

[18] M.C. Morrone and R.A. Owens, "Feature detection from local energy," *Pattern Recognition Letters*, vol. 6, no. 5, pp. 303–313, dec 1987.

[19] Peter Kovesi, "Image Features from Phase Congruency," *Videre*, vol. 1, no. 3, 1999.

[20] Changhan Wang et al., "A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks.," *Conference proceedings : IEEE EMBC*, vol. 2015, pp. 2415–2418, 2015.

[21] Md Murad Hossain, Khalid AlMuhanna, Limin Zhao, Brajesh K Lal, and Siddhartha Sikdar, "Semiautomatic segmentation of atherosclerotic carotid artery wall volume using 3D ultrasound imaging.," *Medical physics*, vol. 42, no. 4, pp. 2029, 2015.

[22] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber, "Training Very Deep Networks," in *Advances in neural information processing systems*, jul 2015, pp. 2377–2385.

[23] E Ukwatta, J Yuan, D Buchanan, B Chiu, J Awad, W Qiu, G Parraga, and a Fenster, "Three-dimensional segmentation of three-dimensional ultrasound carotid atherosclerosis using sparse field level sets.," *Medical physics*, vol. 40, no. May, pp. 052903, 2013.