2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-1179-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/EMBC46164.2021.9631057

Learning-Based Median Nerve Segmentation From Ultrasound Images For Carpal Tunnel Syndrome Evaluation

Mariachiara Di Cosmo, Maria Chiara Fiorentino, Francesca Pia Villani, Gianmarco Sartini, Gianluca Smerilli, Emilio Filippucci, Emanuele Frontoni, Sara Moccia

***Abstract*— Carpal tunnel syndrome (CTS) is the most common entrapment neuropathy. Ultrasound imaging (US) may help to diagnose and assess CTS, through the evaluation of median nerve morphology. To support sonographers, this paper proposes a fully-automatic deep-learning approach to median nerve segmentation from US images. The approach relies on Mask R-CNN, a convolutional neural network that is trained end-to-end. The segmentation head of Mask R-CNN is here evaluated with three different configurations, with the goal of studying the effect of the segmentation-head output resolution on the overall Mask R-CNN segmentation performance. For this study, we collected and annotated a dataset of 151 images acquired in the actual clinical practice from 53 subjects with CTS. To our knowledge, this is the largest dataset in the field in terms of subjects. We achieved a median Dice similarity coefficient equal to 0.931 ($IQR = 0.027$), demonstrating the potentiality of the proposed approach. These results are a promising step towards providing an effective tool for CTS assessment in the actual clinical practice.**

I. INTRODUCTION

Carpal tunnel syndrome (CTS) is the most frequent peripheral neuropathy worldwide [1], with a prevalence of 0.2–4% [2]. CTS is caused by the compression of the median nerve at wrist, as the nerve passes through the carpal tunnel [1]. CTS is encountered with several diseases, including diabetes mellitus, hypothyroidism and rheumatoid arthritis, but cases of idiopathic CTS are also frequent [3]. The diagnosis of CTS is based on clinical history and physical examination [3]. Ultrasound (US) imaging, which is a low-cost and non-invasive procedure, may be used when there is a clinical suspicion of CTS, especially in unclear cases. Specifically, US imaging allows the visualization of the median nerve cross-section, which may be significantly larger in patients with CTS compared to healthy subjects [2].

Performing CTS assessment from US imaging is challenging: US imaging is highly operator dependent and this may lead to low inter-observer reliability [2]. Moreover, as shown in Fig. 1, US images have often low quality, with intensity inhomogeneities, presence of shadows and high noise level. To assist sonographers, the medical-image-analysis community has developed in the years several methods for median nerve segmentation from US images. Hafiane et al. combined a convolutional neural network (CNN) with probabilistic gradient vector flow to delineate the median nerve contour [4]. The method relies on a dataset of US images extracted from 10 videos, each with 500 frames, from 10 patients. In a recent study by Wang et al. [5], a multi-input similarity CNN to track the median nerve was presented. The approach was tested on 100 videos of 6 seconds, each with 180 frames. The videos were collected from 50 patients, which were asked to perform specific wrist motions. In Horng et al. [6], U-Net integrated with a convolutional long short-term memory (LSTM) network was used to process US videos. Ten patients were involved, for a total of 24 videos, each with 420 frames.

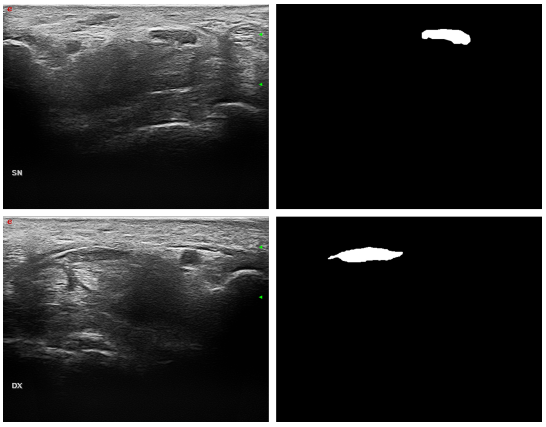

Fig. 1: Sample US images (left column) and associated ground-truth, manually segmented, binary mask of the median nerve (right column). The images are from the same patient. Only the left hand (top row) is affected by CTS.

Despite the promising results, the main limitation of these methods is that they require the manual identification of a region of interest (ROI) around the median nerve. This poses issues relevant to time consumption and inter-clinician variability. To overcome this problem, this work proposes a fully-automatic end-to-end deep-learning approach to accurately detect and segment the median nerve in US images acquired in daily US practice. The contributions of this paper are summarized as follows:

*This work was not supported by any organization

M. Di Cosmo, M.C. Fiorentino and E. Frontoni are with the Department of Information Engineering, Università Politecnica delle Marche, Italy

F.P. Villani is with the Department of Humanities - Languages, Language Liaison, History, Arts, Philosophy, Università di Macerata, Italy, and the Department of Information Engineering, Università Politecnica delle Marche, Italy

G. Sartini, G. Smerilli and E. Filippucci are with the Università Politecnica delle Marche, Italy, and the Rheumatology Unit, Department of Clinical and Molecular Sciences, "Carlo Urbani" Hospital, Jesi (AN), Italy

S. Moccia is with The BioRobotics Institute, Scuola Superiore Sant'Anna and the Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Italy

*Correspondence to M. Di Cosmo: dicosmo.mariachi@gmail.com

TABLE I: Proposed segmentation head. Conv2D: 2D convolution; Conv2DTranspose: transposed 2D convolution.

Operator	Kernel size	No. of filters	Feature size
Conv2D	3x3	256	14x14
Conv2D	3x3	256	14x14
Conv2D	3x3	256	14x14
Conv2D	3x3	256	14x14
Conv2DTranspose	2x2	256	28x28
Conv2DTranspose	2x2	256	56x56
Conv2DTranspose	2x2	256	112x112
Conv2D	1x1	256	112x112

1) Development of an automatic algorithm for median nerve segmentation from US images (Sec. II): Mask R-CNN [7], an end-to-end state-of-art CNN, was studied using different network configurations;
2) Validation in the actual clinical practice (Sec. III): A comprehensive study was conducted collecting in clinical practice 151 images from 53 subjects. This is the largest dataset in terms of patients in the field.

II. METHODS

In this work, we deployed an end-to-end deep-learning algorithm, Mask R-CNN [7], for median nerve semantic segmentation from US images. Mask R-CNN is a CNN made of backbone, Region Proposal Network (RPN), ROIAlign and three heads, for classification, bounding-box regression and segmentation. In this work, architectural changes from the original Mask R-CNN are introduced at the segmentation head to improve output mask resolution.

We use Resnet101 [8] in combination with the Feature Pyramid Network (FPN) [9] as backbone. The FPN allows median nerve detection at multiple scales. The RPN is used to generate proposals, i.e. rectangular regions in the US image with a high probability of containing the median nerve. As in the original implementation, the proposals are predicted starting from anchors, which are here built with 5 different sizes and 3 different scales. The selected proposals are processed by the ROIAlign layer, which adjusts the size of the proposals before they are fed to the Mask R-CNN heads. The classification and regression heads, each made of fully connected layers, predict the proposal class (i.e., median nerve or background) and the bounding box regression values, respectively, thus localizing the median nerve in the image.

The architecture of our segmentation head is shown in Table I and Fig. 2. It is made of four 3x3 convolutional layers with 256 filters. The output of each convolution is activated with the rectified linear unit (ReLU). To recover spatial resolution, upsampling is performed with three transposed convolutions with 256 2x2 filters, ReLU activated. We use three transposed convolution layers, instead of only one as in the original Mask R-CNN, to increase the output resolution and deal with the fragmented and low-contrasted edges of the median nerve. The last layer performs 1x1 convolution and it is activated with the sigmoid function.

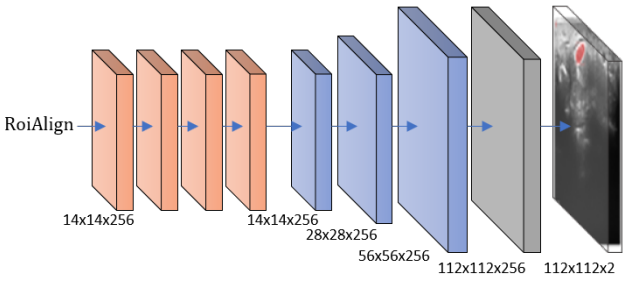

Fig. 2: Visual representation of the segmentation head. The feature-map size is reported. Orange blocks: 2D convolution with 256 3x3 filters, ReLU-activated; Blue blocks: transposed 2D convolution with 256 2x2 filters and stride equal to 2, ReLU-activated; Gray block: 2D convolution with 2 1x1 filters, activated by a sigmoid function.

TABLE II: Performance evaluation metrics. Mean average precision (mAP), Recall (Rec), Precision ($Prec$) are reported.

	mAP	Rec	$Prec$
Mask28	0.896 ± 0.304	0.875 ± 0.261	0.896 ± 0.304
Mask56	0.871 ± 0.341	0.879 ± 0.331	0.848 ± 0.342
Proposed	0.903 ± 0.296	0.903 ± 0.296	0.903 ± 0.296

III. EXPERIMENTAL PROTOCOL

A. Dataset

The US images used in this study were acquired at the Rheumatology Unit of "Carlo Urbani" Hospital, in Jesi (Ancona, Italy) in accordance with the Helsinki Declaration and with the approval of the local ethics committee (Comitato Etico Regione Marche, number 262). All patients signed informed consent. The US assessment was carried out using a MyLab Class C (Esaote SpA, Genoa, Italy) US system working with a 6–18 MHz linear probe, and only transverse scans at the carpal tunnel proximal inlet were taken into account. The dataset consists of 151 US images from 53 patients with an image size equal to 606x468 pixels. Each US image was manually annotated by an expert sonographer with the median nerve contour. The dataset was split over patients, considering 95 images from 36 subjects for training, 25 images from 9 subjects for validation and 31 images from 8 subjects for testing. The US images and their corresponding annotation masks were resized to 512×512 pixels and zero-padded at right-most and bottom-most edges to get squared images with a size multiple of 32, as required by the FPN. In this way, the original aspect ratio was preserved.

B. Training settings

During training, data augmentation was performed on-the-fly by randomly scaling to a value of 80% to 120% of original size and translating of -20% to 20% in both directions, and performing random rotation between (-10°, 10°) and shearing between (-2°, 2°).

Considering the relatively small size of our dataset, transfer learning was used by initializing all layers of the model except for the input layers of the network heads with weights

computed on the COCO (Common Objects in Context) dataset [10]. The training was performed using the Stochastic Gradient Descent as optimizer for 150 epochs with an initial learning rate of 0.001 and momentum of 0.9. A total of 256 anchors per image was used, with varying size (32, 64, 128, 256 and 512) and aspect ratios (1:1, 2:1, 1:2). These values were chosen considering the median nerve section dimension. The ROIAlign resized proposals to a fixed size of 14x14. Hence, the output of the proposed segmentation had a resolution of 112x112.

The network was trained under multi-task cross-entropy loss function combining the loss of classification, localization, and segmentation mask: $L = L_{cls} + L_{bbox} + L_{mask}$, where L_{cls} and L_{bbox} are class and bounding box losses of Faster R-CNN, respectively, and L_{mask} is the mask loss defined in [11].

C. Performance metrics and ablation study

To evaluate the performance in median nerve localization, Precision ($Prec$) and Recall (Rec) were computed as:

$$Prec = \frac{TP}{TP + FP} \tag{1}$$

$$Rec = \frac{TP}{TP + FN} \tag{2}$$

where TP, FP and FN denote the number of true positives, false positives and false negatives, respectively. We considered a TP prediction if the detected bounding box overlapped the bounding box surrounding the ground-truth segmentation for at least 70% and had confidence higher than 0.98. Otherwise, the nerve detection was considered as FP. We considered a FN when no bounding box was predicted at all. Mean Average Precision (mAP), which represents the average of the area under the Recall-Precision curve, was also computed.

The median nerve segmentation performance was measured using the Dice similarity coefficient (DSC), which is defined as:

$$DSC = \frac{2\times \mid A_{gt} \cap A_{mask} \mid}{\mid A_{gt} \mid + \mid A_{mask} \mid} \tag{3}$$

where A_{gt} and A_{mask} are the ground truth and predicted segmentation, respectively. When computing the DSC, only TPs were considered.

As ablation study, we evaluated the effect of having a different number of transposed convolutions in the segmentation head. This was done to assess the effects of an increased resolution of the output of the segmentation head on the overall segmentation performance. The segmentation head was tested with one (Mask28) and two (Mask56) transposed convolutional layers, leading to the output size of the head of 28x28 and 56x56, respectively.

IV. RESULTS

The mAP, Rec and $Prec$ computed on the test set for the proposed model and the ablation-study models are reported in Table II. The best performing model was the proposed one, with $mAP = 0.903 \pm 0.296$, $Rec = 0.903 \pm 0.296$ and $Prec = 0.903 \pm 0.296$. The low standard deviation shows the stability of the model. Specifically, on our 31 test images, the proposed model achieved the highest number of correct detection (28 TP). The median nerve was not identified in only 3 images, all from the same subject. Mask28 produced 28 TP, 3 FN and 2 FP, while Mask56 resulted as the worst performing model, with 27 TP, 4 FN and 4 FP.

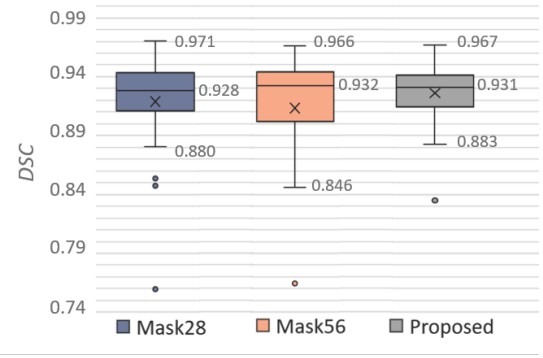

Fig. 3: Boxplot of the Dice similarity coefficient (DSC) computed against the ground-truth manual annotation.

The proposed model achieved a median value of DSC with the narrowest interquartile range (IQR), equal to 0.931 and with $IQR = 0.027$, overcoming both Mask28 and Mask56 as represented in Fig. 3. Sample segmentation results are shown in Fig. 4 to visually compare the results of proposed model with the ground truth and the other models.

V. DISCUSSION

The end-to-end model proposed in this work proved to be a reliable tool for the automatic segmentation of the median nerve in US images. From the experimental results shown in Table II, increasing the output resolution of the segmentation head to 112x112 improved the overall performance. In addition, the proposed model obtained a median DSC (0.931) with the lowest IQR (0.027), thus demonstrating the higher reliability in comparison with Mask28 and Mask56.

The increased output resolution of the segmentation head did not produce any FP, which instead were present both for Mask28 and Mask56, as shown in Fig. 4. The bottom row in Fig. 4 shows an image from the only subject for which the median nerve was not detected by any of the tested models: the poor definition of nerve borders, the contiguity with a vessel of abnormal size and the inhomogeneities of the nerve section could be the causes of the missed detection.

Even if the achieved results are promising, a limitation of this work can be seen in the size of the dataset, even though our dataset is to date the largest in terms of patients in comparison with literature. An original aspect of our work is related to the ability of Mask R-CNN to learn median nerve location and segmentation, without any manual intervention in ROI identification or nerve contour definition. Nevertheless, segmentation performance may be boosted

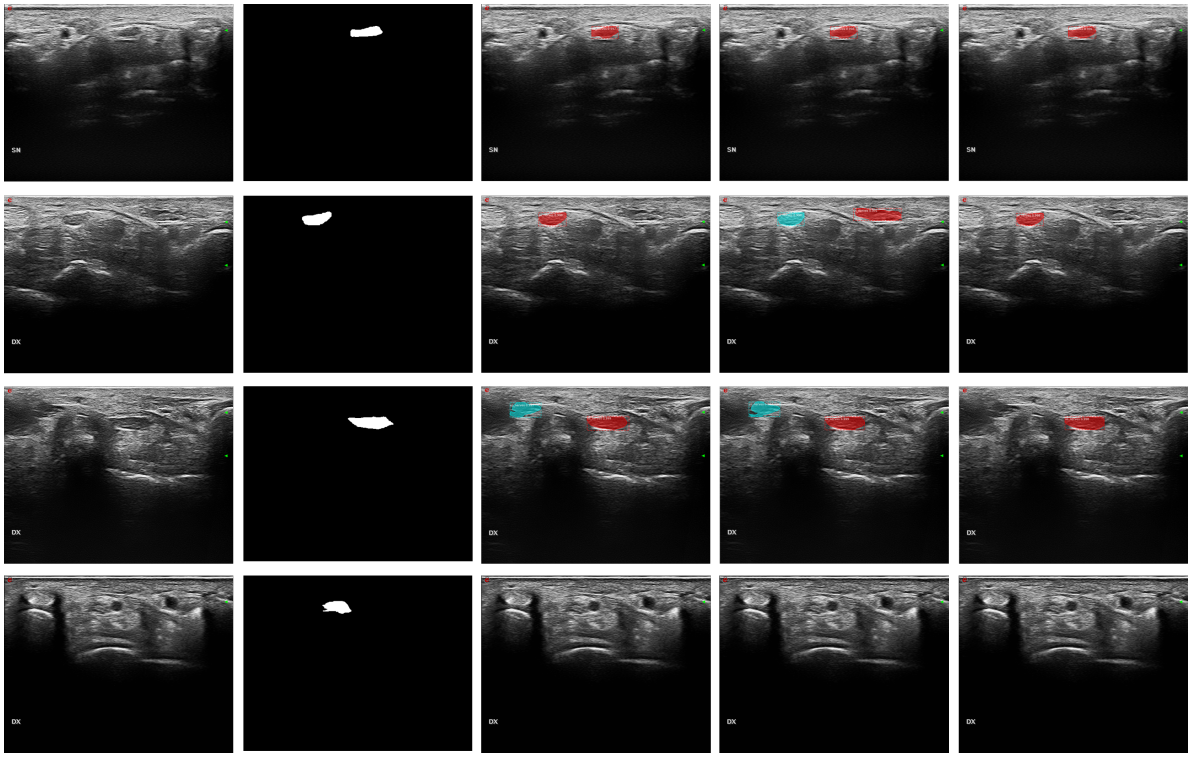

Fig. 4: Visual samples of the segmentation results. From left to right: raw US images, ground truth annotations, predictions obtained with Mask28, Mask56 and the proposed method. First row: all the tested models achieved correct identification of the median nerve; second-third rows: false positives were present for Mask28 and Mask56; last row: sample image from the only patient for which the nerve was not identified by any of the tested models.

further by processing also temporal information naturally encoded in US videos. So far, we focused our analysis on still US images to prevent model overfitting having few patients involved. We are currently working to collect a larger dataset to perform an analysis based on US temporal clips and mimic what is currently done in the actual US practice. This can be done by exploiting spatio-temporal features [12]. Distance-field regression for accurate nerve delineation could be investigated, too, considering the promising results achieved in close fields [13].

VI. CONCLUSION

This work proposed an end-to-end deep-learning approach to median nerve segmentation from US images to help sonographers during carpal tunnel scanning. The obtained results show the potentiality of the proposed approach, opening up to further improvements to fully support real-time US practice.

REFERENCES

[1] L. Padua, D. Coraci, C. Erra, C. Pazzaglia, I. Paolasso, C. Loreti, P. Caliandro, and L. D. Hobson-Webb, "Carpal tunnel syndrome: clinical features, diagnosis, and management," *The Lancet Neurology*, vol. 15, no. 12, p. 1273–1284, 2016.

[2] Y. Yoshii, C. Zhao, and P. C. Amadio, "Recent advances in ultrasound diagnosis of carpal tunnel syndrome," *Diagnostics*, vol. 10, no. 8, p. 596, 2020.

[3] G. Smerilli, A. Di Matteo, E. Cipolletta, S. Carloni, A. Incorvaia, M. Di Carlo, W. Grassi, and E. Filippucci, "Ultrasound assessment of carpal tunnel in rheumatoid arthritis and idiopathic carpal tunnel syndrome," *Clinical Rheumatology*, vol. 40, no. 3, p. 1085–1092, 2020.

[4] A. Hafiane, P. Vieyres, and A. Delbos, "Deep learning with spatiotemporal consistency for nerve segmentation in ultrasound images," *arXiv preprint arXiv:1706.05870*, 2017.

[5] Y.-W. Wang, R.-F. Chang, Y.-S. Horng, and C.-J. Chen, "Mnt-deepsl: Median nerve tracking from carpal tunnel ultrasound images with deep similarity learning and analysis on continuous wrist motions," *Computerized Medical Imaging and Graphics*, vol. 80, p. 101687, 2020.

[6] M.-H. Horng, C.-W. Yang, Y.-N. Sun, and T.-H. Yang, "Deepnerve: A new convolutional neural network for the localization and segmentation of the median nerve in ultrasound image sequences," *Ultrasound in Medicine & Biology*, vol. 46, no. 9, p. 2439–2452, 2020.

[7] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[9] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[10] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," 2014.

[11] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[12] A. Casella, S. Moccia, D. Paladini, E. Frontoni, E. De Momi, and L. S. Mattos, "A shape-constraint adversarial framework with instance-normalized spatio-temporal features for inter-fetal membrane segmentation," *Medical Image Analysis*, p. 102008, 2021.

[13] M. C. Fiorentino, S. Moccia, M. Capparuccini, S. Giamberini, and E. Frontoni, "A regression framework to head-circumference delineation from US fetal images," *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105771, 2021.