A deep learning method for kidney segmentation in 2D ultrasound images

Simão Valente, Pedro Morais, Helena R. Torres, Bruno Oliveira, João Gomes-Fonseca, Buschle L.R., Fritz A., Jorge Correia-Pinto, Estevão Lima, and João L. Vilaça

Abstract—Ultrasound (US) is a medical imaging modality widely used for diagnosis, monitoring, and guidance of surgical procedures. However, the accurate interpretation of US images is a challenging task. Recently, portable 2D US devices enhanced with Artificial intelligence (AI) methods to identify, in real-time, specific organs are widely spreading worldwide. Nevertheless, the number of available methods that effectively work in such devices is still limited. In this work, we evaluate the performance of the U-Net architecture to segment the kidney in 2D US images. To accomplish this task, we studied the possibility of using multiple sliced images extracted from 3D US volumes to achieve a large, variable, and multi-view dataset of 2D images. The proposed methodology was tested with a dataset of 66 3D US volumes, divided in 51 for training, 5 for validation, and 10 for testing. From the volumes, 3792 2D sliced images were extracted. Two experiments were conducted, namely: (i) using the entire database (WWKD); and (ii) using images where the kidney area is > 500 mm² (500KD). As a proof-of-concept, the potential of our strategy was tested in real 2D images (acquired with 2D probes). An average error of 2.88 ± 2.63 mm in the testing dataset was registered. Moreover, satisfactory results were obtained in our initial proof-of-concept using pure 2D images. In short, the proposed method proved, in this preliminary study, its potential interest for clinical practice. Further studies are required to evaluate the real performance of the proposed methodology.

Clinical Relevance—In this work, a deep learning methodology to segment the kidney in 2D US images is presented. It may be a relevant feature to be included in the recent portable US ecosystems, easing the interpretation of image and consequently the clinical analysis.

I. INTRODUCTION

Ultrasound (US) plays a key role in a large variety of clinical applications, such as diagnosis and monitoring of various diseases and surgical procedures [1]. When compared with other imaging modalities, such as X-ray, computed tomography (CT), and magnetic resonance imaging (MRI), the US has the advantage of being more safe, cost-effective, radiation-free, and real-time. Recently, there has been an increasing demand for portable US devices [2], since these devices allow reducing the space occupied in the medical room, allow direct access to US images, promotes medical teaching [3], and due to their ease of use in less developed countries [4]. However, some drawbacks are raised due to its high dependence on the operator’s skills, and to its poor quality and characteristic artifacts, such as shadows, speckle noise, or signal dropout, that hamper the interpretation of the images and may mislead the diagnosis [5][6].

For the reasons listed above, several methods and techniques have been studied and developed by the scientific community for the automatic identification and segmentation of anatomical structures in US images. They have been presented from more traditional and analytical methods of image processing, e.g. using mathematical models, to methods based on Artificial Intelligence (AI) such as Neural Networks and Deep Learning (DL) [7][8]. The use of AI, namely DL, for the identification of anatomical structures, has demonstrated its potential for several areas in clinical practice [7][9][10][11]. Indeed, the potential AI-based methods to automatically analyze US images is currently a strong research topic. Moreover, specific products that use AI to predict the position of the needle in the US image during a surgical procedure are already available in the market [12][13].

In this project, we intend to evaluate the performance of a DL architecture for kidney segmentation. Here, since the acquisition of a large number of US images from multiple views in clinical practice is difficult, we propose to extract multiple 2D US directly from 3D US volumes, ultimately evaluating its performance in real 2D datasets. Therefore, we construct our dataset by extracting slices from 3D volumes, obtaining a large and variable dataset of 2D images.

This paper is organized as follows. In section II, the proposed method is described. In section III, the experiments and their results are presented. In section IV the results are discussed, and in section V the main conclusions of this paper are given.

The authors acknowledge the company KARL STORZ SE & Co. KG for their support of this research. This work was supported by the projects “NORTE-01-0145-FEDER-000045” and “NORTE-01-0145-FEDER-000059”, under the Northern Portugal Regional Operational Programme (NORTE 2020), under the Portugal 2020 Partnership Agreement, through the European Regional Development Fund (FEDER). It was also supported by national funds, through the FCT (Fundaçao para a Ciência e a Tecnologia) and FCT/MCTES in the scope of the project UIDB/05549/2020, UIDP/05549/2020 and LASH-LA/P/0104/2020. The authors also acknowledge FCT, Portugal and the European Social Fund, European Union, for funding support through the “Programa Operacional Capital Humano” (POCH) in the scope of the PhD grants SFRH/BD/136721/2018 (B. Oliveira) and SFRH/BD/136670 (H.R. Torres).

S. Valente is with the 2Ai – School of Technology, IPCA, Barcelos, Portugal, with Algorithmic Center, School of Engineering, University of Minho, Guimarães, Portugal, with Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal, and with ICVS/3B’s PT Government Associate Laboratory, Braga/Guimarães, Portugal, (e-mail: spvalente@ipca.pt).

H. R. Torres and B. Oliveira are with 2Ai – School of Technology, IPCA, Barcelos, Portugal, with Algorithmic Center, School of Engineering, University of Minho, Guimarães, Portugal, with Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal, and with ICVS/3B’s PT Government Associate Laboratory, Braga/Guimarães, Portugal (email: pmorais@ipca.pt, htorres@ipca.pt, boliveira@ipca.pt).

P. Morais, J. Gomes-Fonseca, and J. L. Vilaça are with 2Ai – School of Technology, IPCA, Barcelos, Portugal (email: pmorais@ipca.pt, jlfonseca@ipca.pt, jvilaca@ipca.pt).

Buschle L.R. and Fritz A. are with Karl Storz SE & Co. KG, Tuttinglen, Germany (email: Lukas.buschle@karlstorz.com, andreas.fritz@karlstorz.com.)

J. Correia-Pinto, and E. Lima are with Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal, and with Government Associate Laboratory, ICVS/3B’s-PT, Braga/Guimarães, Portugal (email: jcp@med.uninho.pt, estevao.lima@med.uninho.pt).
II. METHODS

A. Overview

One of the main problems associated with DL methods is the creation of the dataset, since many images with manual annotations is needed to train a neural network. Therefore, in this study, we intend to create a large database by extracting multiple 2D slices directly from 3D volumes. Here, we intend to evaluate if training with sliced US images can be a relevant strategy to augment the available data for 2D segmentation and to achieve accurate segmentation results in real 2D images. Our team has already experience in the development of mathematical models for 3D segmentation of the kidney [14][15][16]. However, the use of 3D imaging for clinical practice is not yet consolidated or at least only used in specialized centers. Therefore, 2D imaging is still a common practice. Additionally, the recent portable devices are still mainly 2D, limiting the integration of the already developed 3D models.

B. Dataset and pre-processing steps

1) 3D US Database

This study was performed using a 66 3D US volume dataset. 19 US 3D volumes (VE9) were acquired with the GE Vivid E9 US scanner, a matrix of array transducers, that works in a range of frequencies between 1.5 and 4.0 MHz, using patient-specific parameters configuration. The remaining 47 US 3D volumes, henceforward called V730, were acquired with the GE Voluson 730 US scanner and an abdominal probe consisting of a hybrid mechanical curved-array transducer with a frequency range of 2-7MHz. The image size across all volumes ranged from 265 x 205 x 182 to 387 x 300 x 243, with an isotropic voxel spacing ranging from 0.486 to 0.9232. The manual delineation of the kidney in the volumes was performed by an experienced observer in the MITK software [17], as described in our previous contribution [14].

The 3D US volumes were randomly divided into three sets: (i) 51 3D US volumes for training, composed of 14 volumes of the VE9 dataset and 37 volumes of the V730 dataset; (ii) 5 3D US volumes for validation, composed of 2 volumes of the VE9 dataset and 3 volumes of the V730 dataset; and (iii) 10 3D US volumes for test, composed for 3 volumes of the VE9 dataset and 7 volumes of the V730 dataset.

2) 2D Database and pre-processing

The 2D dataset was obtained by slicing the 66 3D US volumes separately by training, validation, and testing datasets. This slicing was performed in both sagittal and coronal views, which allows obtaining the standard views of the kidney on 2D US images. Since the kidney is always at the center of the 3D US volumes, and the size of the 3D US volumes in the dataset is variable, the slicing method consisted of extracting slices from ¼ of the volume size to ¾ of the volume size with an incremental step of 5 slices. This procedure resulted in a total of 2924 images in the training dataset, 294 images in the validation dataset, and 574 images in the testing dataset. Due to possible sub-optimal segmentation results in specific slices, all 2D images and corresponding masks were manually analyzed, eliminating those that presented a sub-optimal segmentation. After this post-processing 2166 images, 193 and 358 were included in the training, validation, and testing, respectively, datasets as shown in TABLE I. Finally, as the pixel spacing of the images varies between volumes, a resizing was applied to the images so that they all had the same pixel spacing (0.9232 mm x 0.9232 mm). Afterward, all the images were resized to have 256 x 256 pixels by applying padding/cropping techniques. This methodology is summarized in Figure 1.

C. Kidney Segmentation Method

The 2D segmentation approach relies on Convolutional Neural Network (CNN) model based on U-Net architecture [18] with few adaptations. Similarly, in the U-NET, the network architecture is constituted by the encoding and decoding path. The contraction path includes 5 encoding blocks with 16, 32, 64, 128, and 256 feature maps. Each double convolution of each encoding block is followed by a rectified linear unit (ReLU) used as activation function. A kernel of 3x3 was used for the convolutions. Was used a symmetric expanding path of transposed convolutions with 128, 64, 32, and 16 channels. Afterward, a convolution layer followed by a Sigmoid activation layer was added to the end of the network to achieve the final segmentation.

D. Implementation Details

For the performed experiments in this work (see section III), the network was trained during a maximum of 1000 epochs with a mini-batch size of 20 and using the Adam optimizer with a learning rate of 0.0001. Finally, a He Normal initialization of the model parameter was applied. To guide the network to predict the segmentation, the Dice Loss function was used. To avoid overfitting problems and improve the performance of the networks in 2D acquired US images, data augmentation techniques were implemented, including random brightness adjustments, zooming in, zooming out,

![Figure 1 – Preparation of the 2D dataset.](https://example.com/figure1.png)
shifting up, down, left, and right, rotations, flips, and white noise.

The architecture of this network was implemented in Python 3 language using the Keras deep learning library [19] with TensorFlow as the back end [20]. The computer used in this study contains an Intel® Core™ i7-10700K CPU @ 3.80GHz, 3792 MHz, 8 cores, 16 logical processors, and 32 GB RAM, running on the Windows 10 operating system.

III. VALIDATION

A. Experiments

The generated 2D dataset was used to perform two networks, each one with a different training dataset. The first network was trained with all images of the created 2D dataset (WWKD), where images with and without the kidney are available. This network is henceforth called WWKN. The second network was trained using images that contains the presence of the kidney with an area greater than 500 mm2 (500KD), henceforth called 500KN. This second experiment intends to evaluate the network on a dataset that does only contains images with kidney, removing the influence of images from the surrounding tissues and too small parts of the kidney in the segmentation task. Finally, as proof of concept, the performance of the two networks in the segmentation of 8 2D US images acquired with a 2D US probe was evaluated.

In Table I is presented the amounts of resulting images in each of the datasets created. The WWKD training and validation were used to train the WWKN, while the training and validation data from the 500KD was used to train the 500KN. Test data from the two datasets were used to evaluate the performance of the two implemented networks.

B. Evaluation Metrics

To measure and compare the accuracy of the models, metrics commonly applied for evaluating segmentation methods were selected. Specifically, 8 evaluation metrics were used, namely: Overlap coefficient (Overlap), Dice coefficient (Dice), average Euclidean distance (AverSSD), Hausdorff distance 95th percentile (HfSSSD95), precision, recall, specificity, and accuracy.

C. Results

Table II presents the mean ($\bar{x}$), median (M), and standard deviation (s) results for the two implemented networks with the two validation datasets. When looking for the WWKN and 500KN results, a similar performance was achieved. The better average overlap and Dice of the WWKD on the WWKN can indicate better performance in the identification of the existence or not of a kidney. Overall, the best results are shown on 500KN with the 500KD, showing an average surface distance of 2.88 ± 2.63 mm, with a Dice value of 86.16 ± 11.19%. It should be noted that WWKN correctly identified the existence or not of a kidney in WWKD in 89.66% of cases, against 74.86% of cases in 500KN.

As a final experiment, and as a proof of concept, Figure 2 shows the results of our method (trained with 2D slices extracted from 3D volumes) when tested in real 2D images.

IV. DISCUSSION

In this work, we present a methodology to segment the kidney in 2D US images. Moreover, we also implemented a strategy to create a large dataset of 2D US images, to augment the training data available by directly extracting a set of 2D slices from 3D US volumes.

Two strategies (WWKN and 500KN) were evaluated (using WWKD training and 500KD training datasets, respectively), while validation was performed using two different datasets for both networks (WWKD testing and 500KD testing, respectively). Comparing the results of the two models, we can observe that the 500KN presents a better result in comparison with the WWKN on the validation 500KD dataset. This slight performance improvement is expected since the 500KD dataset was trained with the same type of images (with the presence of kidney) while the WWKD was fed with images that contain kidney and images that do not contain kidney, requiring, therefore, an extra task of identifying the existence or not existence of a kidney in the image. This task had a hit rate of 88.35% against 74.86% in the 500KN with the same images. Looking at the AverSSD results in the four experiments, it is possible to verify a slight increase in the performance of the 500KN for both datasets, demonstrating greater precision in the segmentation of the kidney. On the other hand, the Dice results demonstrate an improvement in the performance of the WWKN compared to the 500KN with the WWKD, demonstrating a greater ability to identify the existence of a kidney. The results in the 8 2D US images acquired with a 2D US probe demonstrated the potential of our strategy, where training with 2D slices extracted from US volumes was performed. In general, the results in the 8 2D US images demonstrate that a network can be trained with sliced 2D US images and obtain good results in the segmentation of 2D US images acquired with a 2D US probe.

![Figure 2](image_url)

Figure 2 Segmentation in images acquired with a 2D US probe. (Green contour – manual label; Blue contour – WWKN; Yellow contour – 500KN).

<table>
<thead>
<tr>
<th>Network</th>
<th>Dataset</th>
<th>Overlap</th>
<th>Dice</th>
<th>AverSSD</th>
<th>HfSSSD95</th>
<th>Precision</th>
<th>Recall</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWK</td>
<td>WWKD</td>
<td>52.91</td>
<td>18.63</td>
<td>30.15</td>
<td>73.58</td>
<td>88.42</td>
<td>80.95</td>
<td>3.32</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td>500KD</td>
<td>52.54</td>
<td>15.95</td>
<td>15.78</td>
<td>84.92</td>
<td>89.91</td>
<td>11.96</td>
<td>3.15</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.52</td>
<td>5.38</td>
<td>7.87</td>
<td>58.79</td>
<td>86.90</td>
<td>43.84</td>
<td>57.77</td>
<td>84.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.71</td>
<td>99.91</td>
<td>6.50</td>
<td>99.29</td>
<td>99.59</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500K</td>
<td>WWKD</td>
<td>52.54</td>
<td>15.95</td>
<td>15.78</td>
<td>84.92</td>
<td>89.91</td>
<td>11.96</td>
<td>3.15</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td>500KD</td>
<td>52.54</td>
<td>15.95</td>
<td>15.78</td>
<td>84.92</td>
<td>89.91</td>
<td>11.96</td>
<td>3.15</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.52</td>
<td>5.38</td>
<td>7.87</td>
<td>58.79</td>
<td>86.90</td>
<td>43.84</td>
<td>57.77</td>
<td>84.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.71</td>
<td>99.91</td>
<td>6.50</td>
<td>99.29</td>
<td>99.59</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE II. EVALUATION RESULTS OF THE WWKN AND 500KN NETWORKS WITH THE WWKD AND 500KD DATASETS.
Comparing the results obtained by our method with other recent DL methods for kidney segmentation in 2D US images, with a dataset acquired with a 2D US probe, it can be considered that our method achieves quite satisfactory results. In [21] a novel CNN model, namely multiscale fusion network of structural feature and detailed (SDFNet) to segment kidney is proposed, obtaining a precision of 96.87%, a Recall of 94.17%, and a Specificity of 99.44%. In [22] is proposed a subsequent boundary distance regression and pixel classification networks to segment the kidney automatically, presenting a Dice coefficient of 93.04%, a Precision of 93.78%, and an AverSSD of 3.03 mm. In this study, we obtained an average Dice of 86.16, a Precision of 90.28%, a Recall of 89.00% a Specificity of 99.63, and an AverSSD of 2.89 mm in the validation dataset. Although a direct comparison cannot be made, the results show that our method using 2D US images sliced from a 3D US volume has similar performance with outer methods and potential to perform a good segmentation task in 2D US images.

Therefore, it seems that the use of 2D sliced images can be a good way to train a deep learning network to learn how to segment 2D US images acquired with a 2D probe. However, further tests with several US images of the kidney, acquired with a 2D probe, are still necessary to validate the real effectiveness of the method. On the other hand, due to the physics of ultrasonic acquisition, slices obtained from a 3D B-mode volume can be different from 2D data in terms of general image properties (e.g. speckle noise and attenuation), so it is still necessary to study whether these differences are significant for network training and if any preprocessing is required to compensate for them.

As future work, we are preparing a 2D dataset of the US kidney with a portable US probe to validate this method and compare the results. Also, it is envisioned to create a new hybrid dataset by merging this new 2D dataset with 2D US images created by slicing a 3D US volume. With this new dataset, further studies will be conducted to understand the added value of 2D sliced US images on the segmentation of 2D US anatomical structures. On the other hand, we are studying a method to slice more images of the same US volume using the slicing method described in [23], which allows the slicing of kidney images from various angles and thus increases the dataset. Although not addressed in the current study, our next goal relies on the integration of the current model directly into the available portable US devices. Indeed, such integration is currently feasible, since specific devices have dedicated APIs to directly stream the available data to external systems. Nevertheless, our system has still some limitations to correctly distinguishing between images with a visible kidney and images without the kidney. To mitigate this issue, a double stage strategy, where a first classification approach is used to detect the presence/absence of the kidney and later perform its segmentation is a potential strategy. Furthermore, and according to our results, the segmentation of small structures is also not straightforward. Indeed, due to our automatic approach to generating 2D slices images, uncommon clinical views were possibly created and integrated into the training. Moreover, the segmentation of small structures purely based on a unique 2D view is extremely challenging, even to manual observers. To solve this problem, we envision developing a real-time method to generate 3D volumes with the segmentation of anatomical structures in real-time through the scanning with a 2D US probe and thus offer a better perception of the US image to the manual observers.

V. CONCLUSION

This work presents a DL method based on the U-Net network to segment the kidney in 2D US images using sliced images obtained from 3D volumes. The results demonstrated a good performance in kidney segmentation of sliced images with a mean average distance error of 2.88 ± 2.63 mm. Moreover, our method showed feasible results in a small proof of concept with real 2D images. In general, this first analysis showed potential to obtain good results, but more tests are needed to evaluate the performance of the method in a large dataset acquired with a 2D US probe.

REFERENCES