Deep Learning-enabled Prostate Segmentation: Large Cohort Evaluation with Inter-Reader Variability Analysis

Yongkai Liu[1], Miao Qi[1,2], Chuthaporn Surawech[1,3], Haoxin Zheng[1], Dan Nguyen[4], Guang Yang[5], Steven Raman[1], and Kyung Sung[1]
[1]Department of Radiological Sciences, University of California, Los Angeles, Los Angeles, CA, United States, [2]Department of Radiology, The First Affiliated Hospital of China Medical University, Shenyang, China, [3]Department of Radiology, King Chulalongkorn Memorial Hospital, Bangkok, Thailand, [4]Department of Radiation Oncology, UT Southwestern Medical Center, Los Angeles, CA, United States, [5]National Heart and Lung Institute, Imperial College London, London, United Kingdom

Synopsis

Whole-prostate gland (WPG) segmentation plays a significant role in prostate volume measurement, treatment, and biopsy planning. This study evaluated a previously developed automatic WPG segmentation, deep attentive neural network (DANN), on a large, continuous patient cohort to test its feasibility in a clinical setting.

Synopsis

Whole-prostate gland (WPG) segmentation plays a significant role in prostate volume measurement, treatment, and biopsy planning. This study evaluated a previously developed automatic WPG segmentation, deep attentive neural network (DANN), on a large, continuous patient cohort to test its feasibility in a clinical setting.

Introduction

Whole-prostate gland (WPG) segmentation plays an important role in prostate volume measurement, biopsy planning, and focal therapy, surgical and radiation planning [1]. However, manual segmentation of WPG is a time-consuming, labor-intensive task and exhibits a high degree of inter-reader variability [2]. Recently, deep learning (DL) [3–6], including deep attentive neural network (DANN) [7], has increasingly been utilized to perform the automated WPG segmentation from MRI images. The evaluation of these methods, to the best of our knowledge, has been performed by relatively small data sizes, ranging from tens to hundreds of MRI scans, limiting their ability to test the DL models in a practical setting. In this study, we evaluate a previously developed DANN [7] using a large, continuous cohort of prostate 3T MRI scans both qualitatively and quantitatively. For qualitative evaluation, the segmentation performance is independently evaluated by two abdominal radiologists via visual grading (Fig 1), where the consistency of the visual grading is tested by measuring the inter-reader agreement. The quantitative evaluation includes segmentation and volume measurement evaluation with manual segmentation as a ground-truth on a small testing set (n=100). The Dice similarity coefficient (DSC) is used to measure the segmentation performance, compared with other baseline DL methods [8], and volume measurement by DANN is compared with the volume measurement by manual segmentation.

Materials and Methods

The study cohort comprised 3,695 MRI scans, acquired on a variety of 3 Tesla MRI scanners from January of 2016 to August of 2020. Axial and coronal T2-weighted (T2W) Turbo spin-echo (TSE) images were used as input to DANN (Fig 3). Of 3,695 3T MRI scans, 335 scans (9%) were used as the training set, and the remaining 3,360 scans (91%) comprised the testing set. Training and testing datasets were randomly chosen from the whole dataset. The testing set included a qualitative segmentation evaluation subset (n=3,210), a quantitative segmentation evaluation subset (n=100), and a volume measurement evaluation set (n=50). Figure 2 shows the data characteristics for each dataset. Training, quantitative, and volume measurement evaluation sets required manual prostate contours as the segmentation reference standard. Figure 3 shows the whole automatic workflow of the WPG segmentation with DANN [7]. First, a DANNcor, responsible for segmenting coronal slices, was adopted to segment the prostate on the two-middle coronal images (9th and 10th slices out of twenty slices) for each MRI scan in the entire testing set. The axial T2W images that contained the prostate gland were then automatically selected by the segmented coronal images. Finally, DANNax was used to perform the WPG segmentation on the selected axial T2W images for each MRI scan in all the testing sets.

Results

Figure 4 shows the result of the qualitative evaluation. For WPG segmentation, 97.9% (n=3,141) and 93.2% (n=2,992) of prostate MRI scans were graded as having acceptable or excellent segmentation performance respectively (Figure 3 (a)). Figure 3 (b) included the confusion matrix to show the interrater variability of the visual grading. Overall, two readers reached a substantial consensus on visual grading in 95.8% of patients (κ =0.74) with an inter-reader grading discrepancy of less than one. Both radiologists unanimously considered 94.6% of segmentation results as acceptable or excellent and graded 91.5% of MRI scans (n=2,861) as having excellent segmentation performance, with only 1.2% of MRI scans (n=39) graded as having unacceptable segmentation performance. Figure 5 shows the result of the quantitative evaluation. For the quantitative evaluation, DANN yielded a DSC of 0.93, significantly higher than the two baseline methods (Deeplab v3+ [9] and UNet [10]) (both p values < 0.05) (Figure 4 (a)). Figure 4 (b) shows the agreement between manual and DANN-enabled volume measurements in the Bland-Altman plot. 48 out of 50 cases (96%) had the volume measurement differences within 95% limits of agreement, indicating that the manual and DANN-enabled volume measurements can be potentially used interchangeably.

Conclusion

Our study showed that a deep learning-based automatic prostate segmentation approach, DANN, could produce prostate segmentation with consistent, sufficient quality when a large, continuous cohort of prostate MRI scans was used for evaluation.

Acknowledgements

No acknowledgement found.

References

1. Garvey B, Türkbey B, Truong H, Bernardo M, Periaswamy S, Choyke PL (2014) Clinical value of prostate segmentation and volume determination on MRI in benign prostatic hyperplasia. Diagnostic Interv Radiol 20:229.

2. Wenger E, Mårtensson J, Noack H, Bodammer NC, Kühn S, Schaefer S, Heinze H-J, Düzel E, Bäckman L, Lindenberger U, others (2014) Comparing manual and automatic segmentation of hippocampal volumes: reliability and validity issues in younger and older brains. Hum Brain Mapp 35:4236–4248.

3. Jin Y, Yang G, Fang Y, Li R, Xu X, Liu Y, Lai X (2021) 3D PBV-Net: An automated prostate MRI data segmentation method. Comput Biol Med 128:104160.

4. Cheng R, Roth HR, Lay NS, Lu L, Turkbey B, Gandler W, McCreedy ES, Pohida TJ, Pinto PA, Choyke PL, others (2017) Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks. J Med imaging 4:41302.

5. Zhu Q, Du B, Turkbey B, Choyke PL, Yan P (2017) Deeply-supervised CNN for prostate segmentation. 2017 Int. Jt. Conf. neural networks. pp 178–184

6. Jia H, Xia Y, Song Y, Cai W, Fulham M, Feng DD (2018) Atlas registration and ensemble deep convolutional neural network-based prostate segmentation using magnetic resonance imaging. Neurocomputing 275:1358–1369.

7. Liu Y, Yang G, Hosseiny M, Azadikhah A, Mirak SA, Miao Q, Raman SS, Sung K (2020) Exploring Uncertainty Measures in Bayesian Deep Attentive Neural Networks for Prostate Zonal Segmentation. IEEE Access 8:151817–151828.

8. Dice LR (1945) Measures of the Amount of Ecologic Association Between Species. Ecology 26:297–302.

9. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). pp 833–851

10. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. Int. Conf. Med. image Comput. Comput. Interv. pp 234–241

Figures

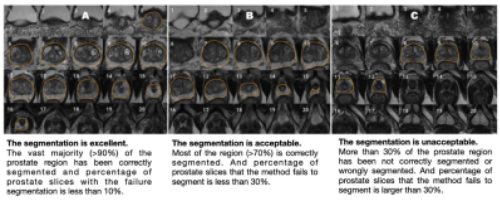

Figure 1: Typical example and description for each visual grade. A, B and C represent a segmentation example with visual grades 3 (excellent), 2 (acceptable), and 1 (unacceptable), respectively. Slice 1-20 represents MRI slices from superior to inferior. Regions encircled by organ boundary are the prostate whole gland.

		Training Dataset	Qualitative Evaluation	Quantitative Evaluation	
			Qualitative segmentation evaluation Dataset	Quantitative segmentation evaluation dataset	Volume measurement evaluation dataset
Number of MRI scans		335	3,210	100	50
MRI scans with different vendor models	Skyra	295	2,806	93	45
	Prisma	10	145	4	3
	Vida	30	259	3	2

Figure 2: Data characteristics in the training, qualitative, and quantitative evaluation.

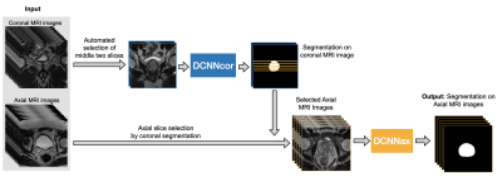

Figure 3: The whole automatic workflow of the WPG segmentation based on DANN. The specific process is as follows: first, DANNcor segmented the WPG on the two middle coronal images (images with the blue border); second, orange lines selected by the prostate segmentation on the coronal images were used to determine the selection of axial slices containing WPG (images with orange border); third, DANNax was performed on the selected axial MRI slices for the segmentation of WPG.

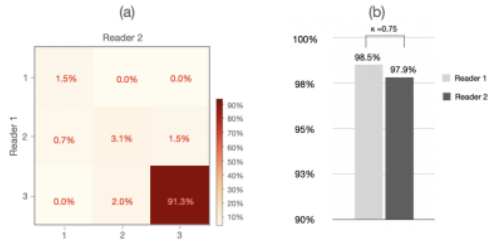

Figure 4: Qualitative evaluation with inter-rater variabilities (n=3,210): (a) Confusion matrices between the visual grades assigned by two readers among all MRI scans; (b) The proportion of segmentation with acceptable or excellent performance evaluated by reader 1 and 2 among all MRI scans. Kappa statistics between the two readers were also provided in the figure.

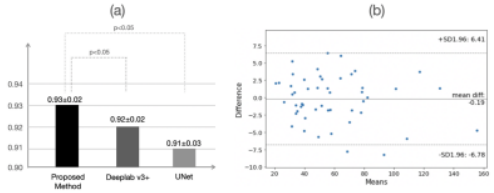

Figure 5: Quantitative evaluations: (a) Bar plot to compare the DSCs achieved by the DCNN and the baseline methods (n=100); (b) Bland–Altman plot to show the agreement between manual and DANN-enabled WPG volume measurements (n=50).