Automated Consensus Contour Building for Prostate MRI

Farzad Khalvati

Abstract— **Inter-observer variability is the lack of agreement among clinicians in contouring a given organ or tumour in a medical image. The variability in medical image contouring is a source of uncertainty in radiation treatment planning. Consensus contour of a given case, which was proposed to reduce the variability, is generated by combining the manually generated contours of several clinicians. However, having access to several clinicians (e.g., radiation oncologists) to generate a consensus contour for one patient is costly. This paper presents an algorithm that automatically generates a consensus contour for a given case using the atlases of different clinicians. The algorithm was applied to prostate MR images of 15 patients manually contoured by 5 clinicians. The automatic consensus contours were compared to manual consensus contours where a median Dice similarity coefficient (DSC) of 88% was achieved.**

I. INTRODUCTION

Prostate cancer is the most commonly diagnosed cancer in North American men, with roughly 23,500 new cases in 2013 in Canada [1] and 238,600 new cases in 2013 in the United States [2]. Furthermore, prostate cancer is the second leading cause of cancer death in both Canadian men with an estimated 3,940 deaths in 2013 [1], and in men in the United States with an estimated 29,720 deaths in 2013 [2]. In the current clinical model, men with positive digital rectal exam and elevated prostate-specific antigen (PSA) require multicore random biopsies for risk stratification. Low-risk patients are usually put under active surveillance to monitor the tumour growth. High-risk patients are provided with treatment, which usually includes surgery or radiation therapy. In diagnosis of prostate cancer, to calculate PSA density, accurate localization and segmentation of the prostate gland in images is required. As well, for delivering treatment via radiotherapy, accurate delineation of prostate gland is required.

Currently, the segmentation (or delineation) of prostate boundaries is performed manually by clinical experts (e.g., radiologists and radiation oncologists) and contouring of a patient's volume dataset is a tedious and tiring task. Nevertheless, the major challenge in prostate delineation is *inter-observer variability*. Inter-observer variability is defined as "the failure by the observer (i.e., clinician) to measure or identify a phenomenon accurately, which results in an error. Sources for this may be due to the observer's missing an abnormality, or to faulty technique resulting in incorrect test measurement, or to misinterpretation of the data" [3].

Inter-observer variability in prostate contouring can affect the PSA density calculation by affecting prostate volume number, potentially negatively influencing the reliability of the active surveillance of prostate cancer patients. In addition, it has been argued that inter-observer variability in anatomic contouring is the most significant contributing factor to uncertainty in radiation treatment planning [4]. This leads to the fact that healthy tissue may receive unnecessary radiation while the cancerous tissue is missed.

Although computed-tomography (CT) imaging is usually used for radiation therapy planning, recent studies have shown promising results for T2-weighted magnetic resonance imaging (T2w-MRI) for prostate cancer monitoring and treatment. The use of MRI could potentially reduce inter-observer variability. For example, it has been reported that the standard error of measurement for prostate volume in T2w-MRI was 4.6 ml where the average prostate volume was 31.9 ml; an inter-observer variability of 14.42% [5]. The standard error of measurement for the same cases for CT and 3D ultrasound was 6.5 ml and 4.9 ml, respectively. However, even with prostate MR imaging, inter-observer variability in contouring remains an obstacle for high quality care.

Consensus contour is the result of combining multiple contours of the same organ (or tumour) from different observers (e.g., radiation oncologists). The goal is to use the crowd wisdom to minimize the error. The Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm [6] uses a collection of contours of a given case to calculate a probabilistic estimate of the true segmentation. It is an Expectation-Maximization algorithm that aims for maximizing the sensitivity and specificity of each input contour with respect to the result (consensus) contour.

A natural way to create a consensus contour would be to have several clinicians manually contour the same organ (e.g., prostate) and then apply the STAPLE algorithm. Nevertheless, in practice, this is highly costly and the healthcare system cannot afford engaging several clinicians (e.g., radiation oncologists) to create contours for a single patient being monitored or going under treatment. In this paper, an algorithm is proposed that automatically generates consensus contours for given cases eliminating the need for several clinicians to manually contour the case. The proposed algorithm is based on atlas-based segmentation (ABS) algorithms; a well-known set of algorithms for medical image segmentation. By automatically generating consensus contour for prostate gland, the error in monitoring and treatment of prostate cancer caused by inter-observer variability is minimized with no extra cost to the healthcare system.

This paper is organized as follows. Section II presents a brief background review of ABS algorithms. Section III describes the proposed automated consensus contour build-

Farzad Khalvati is with the Department of Medical Imaging, Sunnybrook Research Institute, Toronto, Ontario, Canada, M4N 3M5 and the Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1. farzad.khalvati@uwaterloo.ca

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

ing algorithm. Section IV presents the testing methodology. The experimental results and conclusions are presented in Sections V and VI, respectively.

II. Atlas-based segmentation

Atlas-based segmentation (ABS) is a widely used technique for segmenting medical images of different organs (e.g., prostate) [7], [8], [9], [10], [11], [12]. In this method, processed images are stored in a database called atlas along with their manually generated contours. To segment a target image, the atlas images are registered to it and the contour of the best-match image in the atlas is deformed using the registration transformation. In general, there are three approaches to design an ABS algorithm: probabilistic atlas, atlas selection, and multi-atlas approaches. To create a probabilistic atlas [7], the atlas images are usually registered to a manually picked reference image. A mean atlas image is then created by averaging the registered images. Each atlas image is registered against the mean atlas image to produce a deformed contour of the image. The deformed contours are then averaged to generate a probability map of the contours. To segment a target image, the mean atlas image is registered to it and the probability map of the contours is deformed using the registration transform.

The atlas selection approach searches through multiple atlas images to select the one that matches the target image the best. In this method [8], the target image is registered and compared to all atlas images and the best-match atlas contour is selected to be deformed using the corresponding transformation. The multi-atlas approach [12] registers a limited number of atlas images to the target image to create multiple contours, which are then fused using a fusion algorithm to generate the final result. Different fusion algorithms for multi-atlas approaches have been proposed including majority voting and STAPLE algorithms [6].

ABS algorithms mostly rely on the manual contouring provided by the clinicians. The registration algorithms used in ABS transform the atlas images and their manual contours to match the target image and its actual contour, respectively. For a given set of images, each clinician may have their own atlases which contain their manual contouring for the dataset. This means the ABS result for a given target image will heavily depend on the atlases of a specific clinician. The proposed automated consensus contour building algorithm is based on this concept that via an ABS algorithm, it generates segmentation results for a given prostate T2w image using atlases of several clinicians. The generated contours are then combined using STAPLE algorithm to generate the consensus contour for the prostate gland in in T2w image.

III. Proposed automated consensus contour building algorithm

The proposed automated consensus contour building algorithm combines two ABS approaches discussed in Section II, namely atlas selection and multi-atlas. First, the atlases of several clinicians (e.g., radiologists, radiation oncologists) are created by each clinician manually contouring the given dataset. The images and corresponding contours of atlases are stored separately as *user*$_1$ *atlases*, *user*$_2$ *atlases*, ..., *user*$_n$ *atlases*. To generate the consensus contour for a target (unseen) image, first, an atlas selection approach is used to find the best-match atlas image in the atlases of each clinician (user). As discussed in Section II, atlas selection is usually performed by first registering the atlas images to the target image and then comparing the registered atlas images with the target image to find the best-match registered atlas image. Nevertheless, registering all atlas images to the target image is computationally prohibitive and therefore, it limits the number of atlases used. In our approach, we perform the comparison before the registration; the target image is compared to all the atlas images using correlation coefficients and three best-match atlas images are selected. This is performed on each user's atlases separately.

Next, a multi-atlas approach is used to generate a result contour for each user's atlases. First, the three best-match atlas images are registered to the target image. The image registration transformations are then used to deform the contours of the three best-match images extracted from each user's atlases. At this point, for each user's atlases, we have three deformed contours. The STAPLE algorithm is applied to the three deformed contours to generate the candidate contour for each users' atlases. This is basically the consensus contour for each user (*intra-observer consensus contour*). This generates n contours, one for each user's atlases. These intra-observer consensus contours represent each user's input to the final consensus contour. The *inter-observer consensus contour* is then generated by applying the STAPLE algorithm to the n intra-observer consensus contours. Figure 1 illustrates the flow of the algorithm.

IV. Testing Methodology

In the following, details about the testing images, the proposed algorithm, and the performance measure are presented.

A. Image Data

The T2w-MR images with endorectal coil used in this study were derived from an online database[1] provided by Brigham and Women's Hospital. The pulse sequence groups in the DICOM headers of the T2w images were marked as fast relaxation fast spin echo-accelerated (FRFSE-XL). A complete descriptions of the 15 MRI datasets used in this study are provided in Table I. Five radiation oncologists from London Health Sciences Centre, London, Ontario, Canada, manually contoured the MR images for all 15 patients, which included 167 slices.

TABLE I
Description of the prostate T2w MR images

Total studies	Dimensions (pixels)	Resolution (mm)
11	512 × 512	150 × 150 × 3
1	512 × 512	160 × 160 × 3
3	512 × 512	180 × 180 × 3

[1] http://prostateMRimageDatabase.com

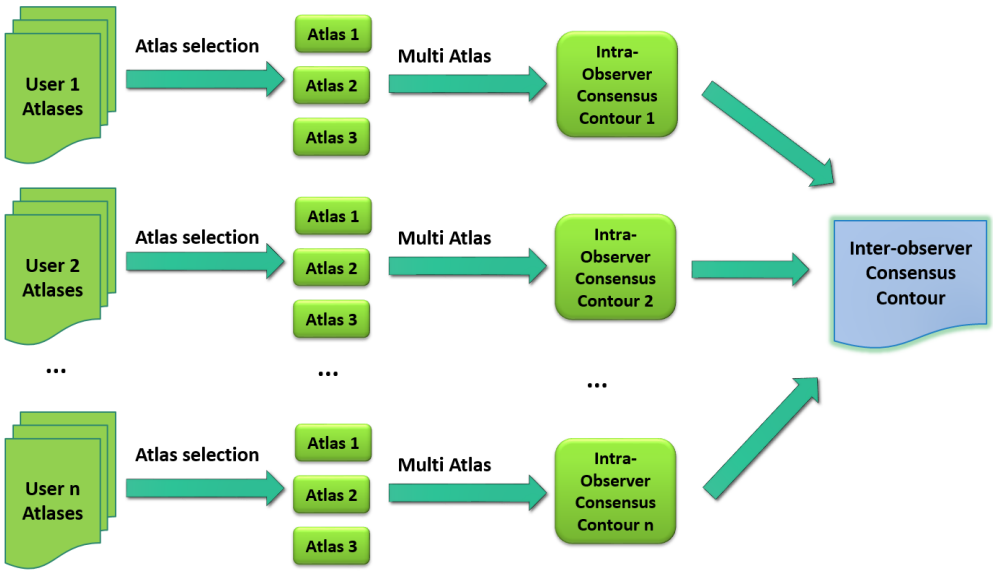

Fig. 1. Automated consensus contour building algorithm

B. Registration Algorithms

The proposed algorithm uses image registration in order to produce the intra-observer consensus contour for each user. To evaluate the performance of the proposed algorithm, we used different registration methods: rigid, affine [13], [14], and nonrigid preceded by affine registration. The nonrigid registration usually captures the local deformations and therefore, it is often beneficial to apply affine registration before applying the nonrigid registration.

For the nonrigid registration, Demon algorithm [15] was used since it is a well-known nonrigid registration algorithm that has been successfully used in registering medical images in different modalities [16], [17]. In our experiments, an open source implementation [18] was used for rigid, affine, and Demon registration algorithms.

C. Slice Accuracy

The accuracy of results was measured by comparing each automatically generated consensus contour with the actual (manual) consensus contour using Dice similarity coefficient (DSC), which is defined as:

$$DSC = \frac{2|C_m \bigcap C_a|}{|C_m| + |C_a|}, \quad (1)$$

where C_m and C_a are the manually and automatically generated contours, respectively. $\bigcap$ represents the shared information in the two binary images.

The manual consensus contour was generated by combining the manual contours of all 5 clinicians for the target image (i.e., the ground-truth contours) via STAPLE algorithm.

V. Experimental Results

The proposed algorithm was implemented in Matlab. Table II summarizes the median and mean DSC values of individual consensus contours (as well as patient accuracy) automatically generated by the proposed algorithm compared to manual consensus contours. The results were obtained via leave-one-patient-out cross-validation approach. It is seen that the best results (88%) are achieved using the rigid registration method within the proposed algorithm[2]. This is interesting since it is usually expected that the nonrigid registration provide a better result. The multi-level consensus contour building approach minimizes the effect of miss-match between the target and atlas images, enabling the rigid registration to produce more accurate results. This is desirable since the processing time for the algorithm with rigid registration is shorter (8 s per contour run in Matlab), which makes it feasible to be used in practice. Figure 2 summarizes the DSC results for all 15 patients.

Figure 3-left shows a sample slice of prostate with the manual and automatic consensus contours. Figure 3-middle and right show the individual manual and atlas result contours for each user, respectively. It is interesting to observe that for this case, the atlas contours (right) are more consistent than the manual contours (middle). The average inter-observer variability among the 5 clinicians for contouring the prostate gland in all 15 MRI datasets was 8.56% .

VI. Conclusions

Inter-observer variability in contouring prostate MR images leads to increased error in segmentation and uncertainty in monitoring and treatment of prostate cancer. The conventional solution requires several clinicians contour the same case. This is costly and not viable in practice. The proposed algorithm in this paper automatically generates the consensus contour for prostate gland in T2w MR images for a given case, eliminating the need for clinicians to actually contour the same case. The proposed algorithm was applied to prostate MR images of 15 patients and a leave-one-out cross-validation was performed. The high accuracy

[2]The rigid registration results were significantly different than that of nonrigid registration ($p \ll 0.01$).

TABLE II

MEDIAN AND MEAN DSC (%) AND PROCESSING TIME (S) VALUES FOR AUTOMATIC CONSENSUS CONTOUR BUILDING

	Median DSC (Slice) (%)	Mean DSC (Slice) (%)	Median DSC (Patient) (%)	Mean DSC (Patient) (%)	Mean Time (Slice) (s)
Rigid	87.65	84.82 ± 9.29	88.01	84.56 ± 4.29	8.37 ± 0.95
Affine	85.82	83.72 ± 9.45	86.86	83.53 ± 4.74	28.19 ± 4.28
Nonrigid	83.36	82.11 ± 7.81	81.57	81.47 ± 3.94	57.21 ± 6.23

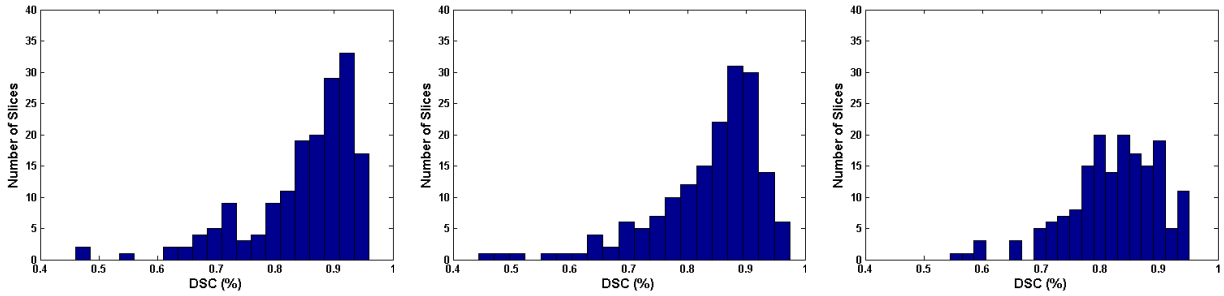

Fig. 2. DSC (%) value comparison between the automatic and manual consensus contours for all 15 patients T2w-MR images using the proposed algorithm with three registration methods. (Left) Rigid registration, (Middle) Affine registration, and (Right) Nonrigid registration.

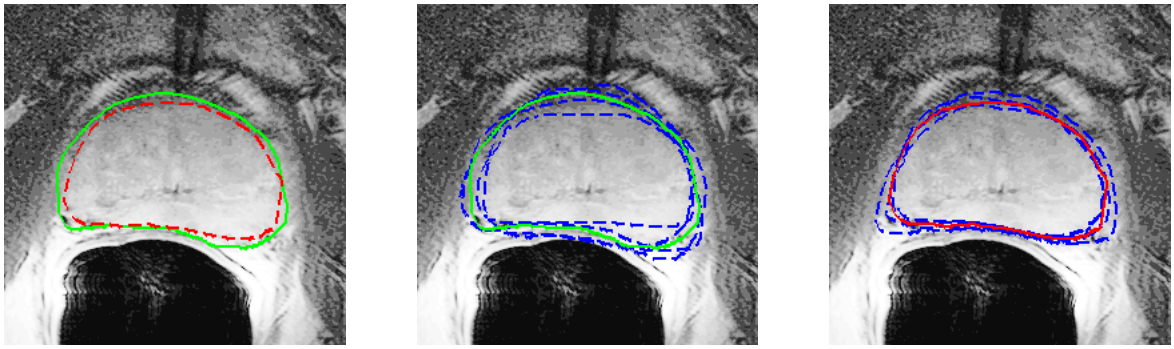

Fig. 3. Automatic consensus contour versus manual consensus contour: Left - solid green and dashed red contours are the manual and automatic consensus contours, respectively. Middle - solid green and dashed blue contours are the manual consensus contour and individual manual contours, respectively. Right - solid red and dashed blue contours are the automatic consensus contour and individual result contours from each user's atlases, respectively.

of the automatic versus manual consensus contours (88%) indicates the potential for the proposed algorithm for future improvement and to be considered as a quality control mechanism for contouring of prostate gland in MR images.

REFERENCES

[1] Canadian Cancer Society. Canadian Cancer Statistics (2013).

[2] American Cancer Society. Cancer Facts and Figures (2013).

[3] U.S. National Library of Medicine, *www.ncbi.nlm.nih.gov/mesh/68015588*.

[4] M. G. Jameson, et al., A review of methods of analysis in contouring studies for radiation oncology. J Med Imaging Radiat Oncol, Vol. 54, pp. 401–410, 2010.

[5] W. L. Smith et al., "Prostate volume contouring: a 3D analysis of segmentation using 3DTRUS, CT, and MR", Int J Radiat Oncol Biol Phys, Vol. 67(4), pp. 1238–47, 2007.

[6] S. K., Warfield, Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging, Vol. 23(7), pp. 903-921, 2004.

[7] S. Martin, et al., Automated segmentation of the prostate in 3D MR images using a probabilistic atlas and a spatially constrained deformable model, Med. Phys., Vol. 37, pp. 1579-1590, 2010.

[8] S. Klein, et al., Automatic segmentation of the prostate in 3d MR images by atlas matching using localized mutual information. Med. Phys, Vol. 35(4), pp. 1407–1417, 2008.

[9] S. Klein, et al., Segmentation of the prostate in MR images by atlas matching, IEEE Conf. Biomed. Imaging, pp. 1300-1303, 2007.

[10] J. Dowling, et al., Automatic atlas-based segmentation of the prostate: A MICCAI 2009 Prostate Segmentation Challenge entry, In: Worskshop in Med. Image Comput. Assist. Interv., pp. 17-24, 2009.

[11] A. Gubern-Mrida, et al., Segmentation of the pectoral muscle in breast MRI using atlas-based approaches, Med. Image Comput. Comput. Assist. Interv., 15(Pt 2), 371–378, 2012.

[12] T. R. Langerak, et al., Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE), IEEE Trans. Med. Imag., Vol. 29, pp. 2000-2008, 2010.

[13] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Med. Image. Anal.,* Vol. 2, pp. 1-36, 1998.

[14] L. G. Brown, "A survey of image registration techniques," *ACM. Computing. Surveys.,* Vol. 24, pp. 325-376, 1992.

[15] J. P. Thirion, "Image matching as a diffusion process: an analogy with Maxwell's demons," *Med. Image. Anal.,* Vol. 196, pp. 243-260, 1998.

[16] A. Guimond, et al., "Three-dimensional multimodal brain warping using the demons algorithm and adaptive intensity corrections, *IEEE Trans. Med. Imag.,* Vol. 20(1), pp. 58-69, 2001.

[17] H. Wang, et al., "Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy", *Phys. Med. Biol.*. Vol. 50(12), pp. 2887-2905, 2005.

[18] D J. Kroon, "Multimodality nonrigid demon algorithm image registration", MatlabCentral, *www.mathworks.com/matlabcentral/fileexchange/21451-multimodality-non-rigid-demon-algorithm-image-registration*