Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information

Stefan Klein[a)]
*Image Sciences Institute, University Medical Center Utrecht, Q0S.459, P.O. Box 85500,
3508 GA Utrecht, The Netherlands*

Uulke A. van der Heide, Irene M. Lips, and Marco van Vulpen
*Department of Radiotherapy, University Medical Center Utrecht, Q00.118, P.O. Box 85500,
3508 GA Utrecht, The Netherlands*

Marius Staring and Josien P. W. Pluim
*Image Sciences Institute, University Medical Center Utrecht, Q0S.459, P.O. Box 85500,
3508 GA Utrecht, The Netherlands*

(Received 25 September 2007; revised 18 January 2008; accepted for publication 18 January 2008; published 14 March 2008)

An automatic method for delineating the prostate (including the seminal vesicles) in three-dimensional magnetic resonance scans is presented. The method is based on nonrigid registration of a set of prelabeled atlas images. Each atlas image is nonrigidly registered with the target patient image. Subsequently, the deformed atlas label images are fused to yield a single segmentation of the patient image. The proposed method is evaluated on 50 clinical scans, which were manually segmented by three experts. The Dice similarity coefficient (DSC) is used to quantify the overlap between the automatic and manual segmentations. We investigate the impact of several factors on the performance of the segmentation method. For the registration, two similarity measures are compared: Mutual information and a localized version of mutual information. The latter turns out to be superior (median $\Delta\mathrm{DSC} \approx 0.02$, $p < 0.01$ with a paired two-sided Wilcoxon test) and comes at no added computational cost, thanks to the use of a novel stochastic optimization scheme. For the atlas fusion step we consider a majority voting rule and the "simultaneous truth and performance level estimation" algorithm, both with and without a preceding atlas selection stage. The differences between the various fusion methods appear to be small and mostly not statistically significant ($p > 0.05$). To assess the influence of the atlas composition, two atlas sets are compared. The first set consists of 38 scans of healthy volunteers. The second set is constructed by a leave-one-out approach using the 50 clinical scans that are used for evaluation. The second atlas set gives substantially better performance ($\Delta\mathrm{DSC} = 0.04, p < 0.01$), stressing the importance of a careful atlas definition. With the best settings, a median DSC of around 0.85 is achieved, which is close to the median interobserver DSC of 0.87. The segmentation quality is especially good at the prostate-rectum interface, where the segmentation error remains below 1 mm in 50% of the cases and below 1.5 mm in 75% of the cases. © *2008 American Association of Physicists in Medicine*. [DOI: 10.1118/1.2842076]

Key words: prostate, segmentation, atlas, nonrigid registration, localized mutual information

I. INTRODUCTION

Prostate cancer treatment by radiation therapy requires an accurate localization of the prostate: Neighboring tissue (rectum and bladder) should be spared, while the tumor should receive a prescribed dose. For the treatment planning, computed tomography (CT) images are primarily used, but increasingly magnetic resonance (MR) images are added, because of their soft-tissue contrast.[1,2] Several studies[1,3] have demonstrated that the additional use of MR images for prostate delineation leads to a reduced interobserver variation and a smaller estimated prostate volume. In current practice at our hospital a manual delineation of the prostate is made, based on the CT and MR scans, which is a labor-intensive task and requires training. Therefore, automating this process is desired.

Figure 1 shows two example MR slices together with their manual delineations. The MR protocol, a balanced steady-state free precession (bSSFP) sequence, was optimized for visibility of the prostate and rectum. An extensive review of the prostate's anatomy visible on MR images can be found in Ref. 2. The major components of the prostate are the central gland, the peripheral zone, and the seminal vesicles, each having different appearances on the bSSFP MR scans. The shape and size of the seminal vesicles vary heavily among people. The central gland and the peripheral zone together have the size of a walnut (around 25 ml) in healthy subjects. Prostate cancer develops most frequently in men over 50. With increasing age, a large group of men also suffers from benign prostate hypertrophy (BPH), which can result in substantial growth of the central gland.

0094-2405/2008/35(4)/1407/11/$23.00

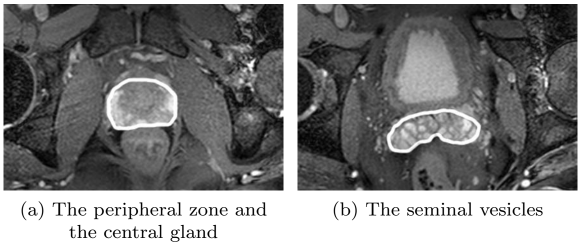

(a) The peripheral zone and the central gland
(b) The seminal vesicles

FIG. 1. Two example MR slices, zoomed in on the region of interest, with manually delineated prostate (white line).

Challenges for automatic segmentation of the prostate in MR images include the presence of imaging artifacts due to air in the rectum and inhomogeneities of the magnetic field, the large anatomical variability between subjects, the differences in rectum and bladder filling, and the lack of a normalized "Hounsfield" unit for MR. Four examples of imaging artifacts, taken from the clinical test data described in Sec. III A 2, are shown in Fig. 2.

Recent surveys of the literature on automatic segmentation of the prostate can be found in Refs. 4 and 5. Existing work has mainly focused on statistical model based approaches. In Ref. 6, a pseudo-three-dimensional (3D) active shape model is used to segment the prostate without seminal vesicles in MR images. In Ref. 4, a method is proposed that combines a statistical model for the prostate with region-growing methods for the rectum and the bladder. The seminal vesicles are not included and manual initialization is required. In Ref. 7 a method to segment pelvic CT images is presented that uses *intrasubject* nonrigid registration of a manually segmented planning scan.

In this article, we propose a fully automatic method to delineate the prostate including the seminal vesicles in 3D MR scans. The method is based on *intersubject* registration of atlas images. The atlas consists of a set of manually labeled MR images from multiple individuals. Using a nonrigid registration algorithm, all atlas images are matched to the patient's MR image that is to be segmented. The deformed manual segmentations of the atlas images are combined into a single segmentation of the patient's image (label fusion). Multiple atlas images are used, instead of a single image, to account for the large anatomical variability between subjects and for the differences in bladder and rectum filling. Multi-atlas based segmentation methods have given promising results in other applications.[8]

Three factors that may influence the performance are investigated: The similarity measure used by the registration process, the atlas label fusion method, and the composition of the atlas set. In Sec. II A, the two similarity measures are described: Mutual information and a localized version of mutual information. For the latter we introduce a novel stochastic optimization method. The atlas label fusion methods are treated in Sec. II B. The atlas sets that are used in the experiments are described in Sec. III.

The proposed method is evaluated on 50 clinical scans. To determine the ground truth, each scan was manually segmented by three human experts. The Dice similarity coefficient (Sec. III C) is used to quantify the overlap between automatic and manual segmentations. The spatial distribution of the segmentation errors is visualized using a spherical coordinate mapping of the prostate boundary. Section IV presents the results of the experiments. First, the impact of the three factors mentioned above is explored. Subsequently, the accuracy of the automatic segmentation obtained with the optimum settings is compared to the interobserver variability. Recommendations for future work are given in Sec. V and the article is concluded in Sec. VI.

II. METHOD

The patient's image to be segmented is denoted by $P(\boldsymbol{x})$. The goal of the automatic segmentation method is to produce a binary label image $L(\boldsymbol{x})$ that accurately defines the prostate of the patient.

The proposed segmentation method follows the general scheme of multi-atlas based segmentation methods, see, for example, Ref. 8. A set of M accurately labeled images, which serve as an atlas, is assumed to be available. The ith image in this atlas set is referred to as $A_i(\boldsymbol{x})$. The corresponding label image, created by a human expert, is called $L_i(\boldsymbol{x})$. The segmentation method consists of two stages: (1) Registration and (2) label image fusion. In the registration stage, each atlas image A_i is matched to the patient image P, using a nonrigid registration algorithm. The resulting coordinate transformations are applied to the label images L_i. In the label image fusion stage, the deformed label images are combined into a single segmentation L of the target patient im-

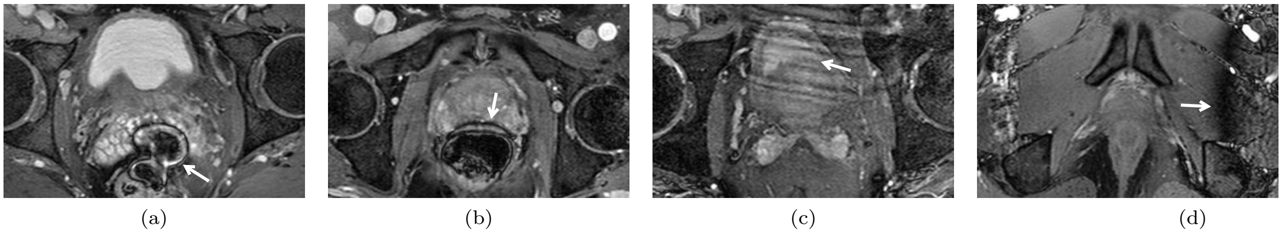


FIG. 2. Four examples of imaging artifacts, marked by white arrows: (a) and (b) Susceptibility artifacts due to air in the rectum, which manifest themselves as black lines not corresponding to tissue boundaries, (c) low-contrast prostate-bladder boundary combined with a streaking artifact, (d) large intensity inhomogeneity.

age. Note that in all steps the images are treated as 3D volumes, rather than processing them on a two-dimensional slice-by-slice basis.

II.A. Registration

In the registration stage, each atlas image A_i is matched to the patient image P. A coordinate transformation $T_i(\boldsymbol{x})$ is estimated that maximizes the similarity of P and the deformed atlas $A_i \circ T_i$ [the symbol $\circ$ represents function composition: $(A_i \circ T_i)(\boldsymbol{x}) = A_i(T_i(\boldsymbol{x}))$]. The registration is performed in two steps. First, rough alignment of the two images is achieved by a rigid registration. After that a nonrigid registration is performed, using a coordinate transformation that is parameterized by cubic B-splines.[9] The parameters that describe the transformation are represented by the vector $\boldsymbol{\mu}$.

An important aspect of the registration method is the choice of the similarity measure. We compare two similarity measures: Mutual information (MI)[10,11] and localized mutual information (LMI).[12,13] The mutual information of two d-dimensional images $I(\boldsymbol{x}), J(\boldsymbol{x}) : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ is defined as follows:

$$MI(I,J;\Omega) = \sum_k \sum_m p_{IJ}(k,m) \log \frac{p_{IJ}(k,m)}{p_I(k) p_J(m)}, \tag{1}$$

where p_I and p_J denote the discrete marginal intensity probabilities of I and J, respectively, and p_{IJ} represents the discrete joint intensity probability. The intensity probabilities are estimated from a discrete set of intensity pairs $[I(\boldsymbol{x}_i), J(\boldsymbol{x}_i)]$, where the coordinates $\boldsymbol{x}_i$ are sampled from the continuous image domain Ω. A common choice is to use all voxel locations, or a uniformly sampled subset of those. An important assumption of MI is that the true intensity probabilities do not vary over Ω. This assumption is often violated in MR scans, due to the presence of magnetic field inhomogeneities. Therefore, it may be better to evaluate the mutual information on multiple subregions, each having a more stationary intensity distribution. Adding the resulting mutual information values of all subregions gives us the localized mutual information LMI[12,13]

$$LMI(I,J;\Omega) = \frac{1}{N} \sum_{\boldsymbol{x}_j \in \Omega} MI(I,J;\mathcal{N}(\boldsymbol{x}_j)). \tag{2}$$

In this equation $\mathcal{N}(\boldsymbol{x}_j) \subseteq \Omega$ represents a spatial neighborhood centered on $\boldsymbol{x}_j$. The number of neighborhoods is denoted by N. The neighborhood center coordinates $\boldsymbol{x}_j$ are samples from Ω. We may choose them to be all voxel locations, or some subset of those. The neighborhoods $\mathcal{N}(\boldsymbol{x}_j)$ must be chosen large enough to allow for a reliable estimation of the intensity probabilities, but small enough to ensure that the influence of the inhomogeneities is negligible. We considered cubic regions of 25^3, 50^3, and 100^3 mm, and compared their performance in 36 registrations on a subset of the data described in Secs. III A 1 and III A 2. Six scans of the first data set were registered to six scans of the second data set. The best results in terms of the prostate overlap after registration, see Sec. III C, were obtained with the 50×50×50 mm region. This setting is used in all experiments that are described in this article.

For maximization of the similarity measure we employ an iterative optimization routine, called stochastic gradient descent. The parameters $\boldsymbol{\mu}$ that describe the transformation are updated in each iteration k by taking a step in the direction of the derivative of the similarity measure with respect to $\boldsymbol{\mu}$. In Ref. 14 it is demonstrated for MI that convergence to the solution is still achieved when the derivative is approximated using only a very small number P of randomly sampled intensity pairs. Two important conditions for this are that new samples are selected in every iteration and that the step size a_k is a slowly decaying function of the iteration number k,

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k \tilde{\boldsymbol{g}}_k, \tag{3}$$

$$a_k = a/(k+A)^{\alpha}, \tag{4}$$

where $\tilde{\boldsymbol{g}}_k$ represents the approximated derivative of MI, and $a > 0$, $A \geqslant 1$, and $0 \leqslant \alpha \leqslant 1$ are user-defined constants. For LMI we can use the same strategy and even extend it by using a small set of neighborhoods, randomly selected in every iteration. We use $N=1$, in other words, LMI is implemented by computing MI using P intensity pairs, sampled from a 50×50×50 mm neighborhood that is randomly selected in every iteration. This approach results in equal computational costs per iteration for MI and LMI, provided that the same number of intensity pairs P are sampled to estimate p_{IJ}. The stochastic optimization procedure that we use is an important difference to Refs. 12 and 13.

The registration algorithm was integrated in `elastix` (`www.isi.uu.nl/Elastix`), a publicly available package for medical image registration, developed by the authors. The mutual information MI is implemented according to Ref. 15, using a joint histogram size of 32×32 and cubic B-spline Parzen windows. The number of samples randomly selected in each iteration is set to $P=2000$. A four-level multiresolution scheme is employed in both the rigid and the nonrigid registration step. Gaussian smoothing is applied to the image data using a standard deviation of 4, 2, 1, and 0.5 voxels in the four respective resolutions. The nonrigid registrations are performed using a B-spline control point spacing of 64, 32, 16, and 8 mm in all directions, for the four respective resolutions. Per resolution, 2000 iterations are performed. The step size sequence in Eq. (4) is denned by $a=2000$, $A=200$, and $\alpha=0.6$. The above described settings were determined by trial-and-error experiments on two image pairs, randomly selected from the data set described in Sec. III A 1.

II.B. Label image fusion

The registration stage yields a set of transformations T_i, which can be applied to the atlas label images L_i, resulting in a set of deformed label images $L_i \circ T_i$, $i=1,\ldots,M$. These must be combined into a single segmentation of the patient's image. For this purpose we consider majority voting (VOTE) and "simultaneous truth and performance level estimation" (STAPLE), explained in Secs. II B 2 and II B 3, respectively.

Both methods can be combined with a preceding atlas selection stage, which is described in Sec. II B 1. In the experiments we compare VOTE and STAPLE, both with and without the atlas selection procedure.

The voxels of the atlas label images L_i take discrete values $c \in \mathcal{C}$, each corresponding to a certain tissue type (class), with $\mathcal{C}$ the set of classes. For example, "1" represents prostate tissue, "2" represents the bladder, and "0" everything else. Although the aim of our work is segmentation of only the prostate, the label fusion procedures VOTE and STAPLE may benefit from additionally labeled tissue types in the atlas. In the experiments this aspect is investigated.

II.B.1. Atlas selection

Instead of using all deformed label images we can make a selection of atlas scans and use only their associated deformed label images. The selection is based on the similarity of the patient image P and the deformed atlas images $A_i \circ T_i$. As in Ref. 8, we measure the similarity after registration by the normalized mutual information (NMI).[16] Let us define the ratio r_i,

$$r_i = \frac{NMI(P, A_i \circ T_i; \Omega)}{\max_j NMI(P, A_j \circ T_j; \Omega)}. \tag{5}$$

An atlas A_i is selected if it satisfies $r_i \geqslant \varphi$, where $0 \leqslant \varphi \leqslant 1$ is a tunable parameter. A value of 0 means that all atlas scans are included in the selection. A value of 1 implies that only the atlas scan with the highest similarity measure is used. The settings $\varphi = 0$ and $\varphi = 1$ correspond to the "MUL" and "SIM" methods, respectively, investigated in Ref. 8. In Sec. IV, we present results for a range of φ.

The set of atlas image indices selected in this stage is called $\mathcal{A}_P$. The subscript indicates that this set can be different for each patient image.

II.B.2. Majority voting (VOTE)

To combine the deformed segmentations of the selected atlas images into a single segmentation $L(\boldsymbol{x})$, majority voting is the most straightforward method. We consider a somewhat more general, weighted version, defined by the following two equations:

$$\ell_c(\boldsymbol{x}) = \frac{\sum_{i \in \mathcal{A}_P} w_i \cdot \delta[c, (L_i \circ T_i)(\boldsymbol{x})]}{\sum_{i \in \mathcal{A}_P} w_i}, \quad \forall\, c \in \mathcal{C}, \tag{6}$$

$$L(\boldsymbol{x}) = \arg\max_{c \in \mathcal{C}} \ell_c(\boldsymbol{x}), \tag{7}$$

where $\ell_c(\boldsymbol{x})$ denotes the probability of class c at $\boldsymbol{x}$, $\delta[\cdot]$ is the Kronecker delta function, and w_i are scalar weighting factors. Equation (7) selects the class with the highest probability as the final label. Setting $w_i = 1$ for all i yields the common majority voting procedure. By using $w_i = r_i$ more weight is assigned to atlas scans that match well to the patient image. Both approaches are tested in Sec. IV.

II.B.3. Simultaneous truth and performance level estimation (STAPLE)

The STAPLE algorithm[17,18] treats the label image fusion as a maximum-likelihood problem, which is solved using an expectation-maximization procedure. Intuitively, the method is based on the following two observations: (1) If the patient segmentation L is known, the accuracy (reliability) of each deformed label image $L_i \circ T_i$ can be computed in terms of its specificity and sensitivity, and (2) if the specificity and sensitivity values of all deformed label images are known, a better estimate of L can be generated. In Ref. 18 it is demonstrated that the STAPLE algorithm gives better results than VOTE, when used for atlas-based segmentation of bee brains.

We run the STAPLE procedure using the *disputed* voxels only, i.e., the voxels where $(L_i \circ T_i)(\boldsymbol{x}) \neq (L_j \circ T_j)(\boldsymbol{x})$ for at least one combination of $i, j \in \mathcal{A}_P$. When the deformed label images are reasonably similar to each other, the disputed voxels lie on a narrow band around the prostate border. The STAPLE algorithm needs to be initialized with a probabilistic segmentation of each class. The probabilistic segmentation ℓ_c that results from VOTE, see Eq. (6), is a reasonable choice for this. The choice of w_i in Eq. (6) may influence the final STAPLE result, although the effect can be expected to be small, since the VOTE procedure serves here as an initialization only. In Sec. IV both $w_i = 1$ and $w_i = r_i$ are tested.

Note that, if all deformed label images indicate an over- or undersegmentation of the prostate, the final label image L will also be an over- or undersegmentation. This happens regardless of the label image fusion method (VOTE or STAPLE) and is not affected by the decision to use only the disputed voxels.

III. EXPERIMENTS

Two data sets are available for the evaluation. The first set consists of 38 scans, originating from healthy volunteers. The second set consists of 50 clinical scans from prostate cancer patients.

For the experimental evaluation of the proposed segmentation method, an atlas needs to be defined. The composition of the atlas may have a large impact on the quality of the segmentations. The atlas should contain enough anatomical variation, such that for every target patient image a few atlas images are present that are reasonably similar to the patient image, allowing for successful registration. If the images are of very high quality, the diversity of the atlas may not be so important anymore, since the registration algorithm would match any pair of images successfully. We evaluate the influence of the atlas composition in our application by performing two types of experiments. In the first experiment the volunteer data set serves as an atlas and the clinical data set serves as a test set. The second experiment is a leave-one-out test, using only the patient data.

All experiments are performed both with MI and LMI as the similarity measure for registration. Also, the various atlas label fusion procedures described in Sec. II B are tested. The

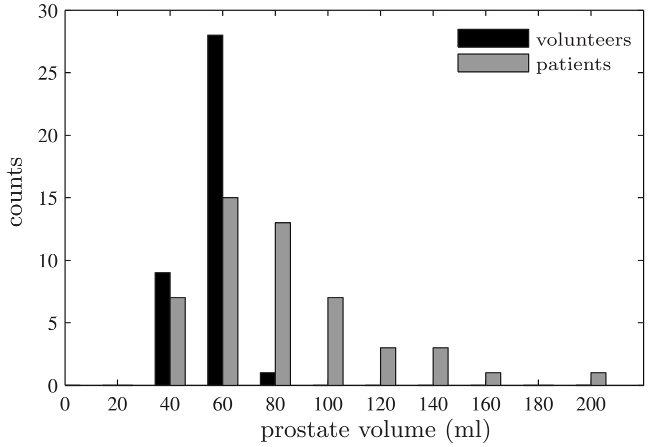


FIG. 3. Histogram of prostate volumes. The mean volume (± st.dev.) is 52 ± 6 ml for the volunteers and 82 ± 36 ml for the patients.

results are evaluated by comparing the automatically generated segmentations with manual segmentations.

III.A. Data

III.A.1. Volunteer data

The volunteer data set consists of 38 MR scans, acquired with a Philips 3T scanner (Gyroscan NT Intera, Philips Medical Systems, Best, The Netherlands) using a flex-M coil and a balanced steady-state free precession (bSSFP) sequence with fat suppression. The scans originate from eight healthy volunteers (age 42–51 years, mean 47) and were made in the context of another study. Seven volunteers were scanned five times, one volunteer was scanned three times. The time between two scans was at least one day, and the volunteers were asked to try to vary the content of their rectum and bladder, to get as much variety between the scans as possible. The scans have a dimension of $512 \times 512 \times 90$ voxels of size $0.49 \times 0.49 \times 1.0$ mm. Manual segmentations are available for each scan. The segmentations were made by an experienced observer and approved by a radiation oncologist (observer A, see Sec. III A 2). Note that the seminal vesicles are considered part of the prostate. Besides the prostate, the bladder and the rectum were also labeled. The distribution of prostate volumes is visualized in Fig. 3 by the black bars of the histogram.

III.A.2. Patient data

The 50 clinical scans were acquired using the same protocol as the scans in the volunteer data set and originate from 50 prostate cancer patients (age 51–79 years, mean 69), who were scheduled for external beam radiation therapy. The patients did not have any loco-regional or distant metastases. For 35 patients the disease status was $T_{3,4}N_0M_0$. The rest was classified as $T_{1,2}N_0M_0$. In each scan the prostate was segmented by three observers. Observer A is a radiation oncologist and has the most experience (ten years) of the three observers. Observer B is a resident radiation oncologist and observer C is a medical physicist specialized in the field of prostate radiotherapy. We constructed an additional "gold standard" L^G by combining the three segmentations L^A, L^B, and L^c using majority voting, with equal weights w_i. Note that only the prostate was delineated in the patient data. Bladder and rectum were not labeled. The distribution of prostate volumes as defined by L^G is shown in Fig. 3 by the gray bars. Clearly, a much larger range of prostate volumes is present in the patient data set than in the volunteer data set. It is well known that men in the age group of the patients often suffer from benign prostate hypertrophy (BPH). This can result in a substantial increase of the prostate volume.

III.B. Experiment description

III.B.1. Experiment I

In Experiment I the volunteer data set serves as an atlas and the clinical data set serves as a test set. For the label image fusion algorithms, VOTE and STAPLE, two choices of $\mathcal{C}$ (see Sec. II B) are considered: $\mathcal{C}=\{\text{background, prostate}\}$ and $\mathcal{C}=\{\text{background}, \text{prostate}, \text{rectum}, \text{bladder}\}$, where "background" is defined as anything that does not belong to one of the other classes. The resulting label image fusion methods are referred to as VOTE2, VOTE4, STAPLE2, and STAPLE4, where the number indicates the number of classes in $\mathcal{C}$. Note that the rectum and bladder segmentations that come as a by-product from VOTE4 and STAPLE4 are not of our interest. We only assess the quality of the prostate delineation. As mentioned in Sec. II B, the presence in the atlas of additionally labeled tissue types besides prostate may improve the segmentation of the prostate in the target patient image.

During the registration of the atlas images to the patient images the similarity measure (MI or LMI) is evaluated on a region of interest Ω. A rectangular region of interest of $271 \times 333 \times 86$ voxels was manually selected for this purpose, roughly encompassing the prostate, bladder and rectum in all scans. For atlas selection, see Eq. (5), the same Ω is used.

III.B.2. Experiment II

The second experiment is a leave-one-out test, using only the patient data. For each patient the atlas set thus consists of the 49 remaining patients. The gold standard labels L^G are used as atlas label images. Only VOTE2 and STAPLE2 are considered in Experiment II, since no manual segmentations of the rectum and bladder are available for the patient data. For Ω the same definition is used as in Experiment I.

It may be expected that the results of Experiment II are better than those of Experiment I, since the atlas contains more anatomical variation, as shown in Fig. 3. We evaluate the relative impact of this difference in atlas composition, compared to other factors that influence the performance of the automatic segmentation method.

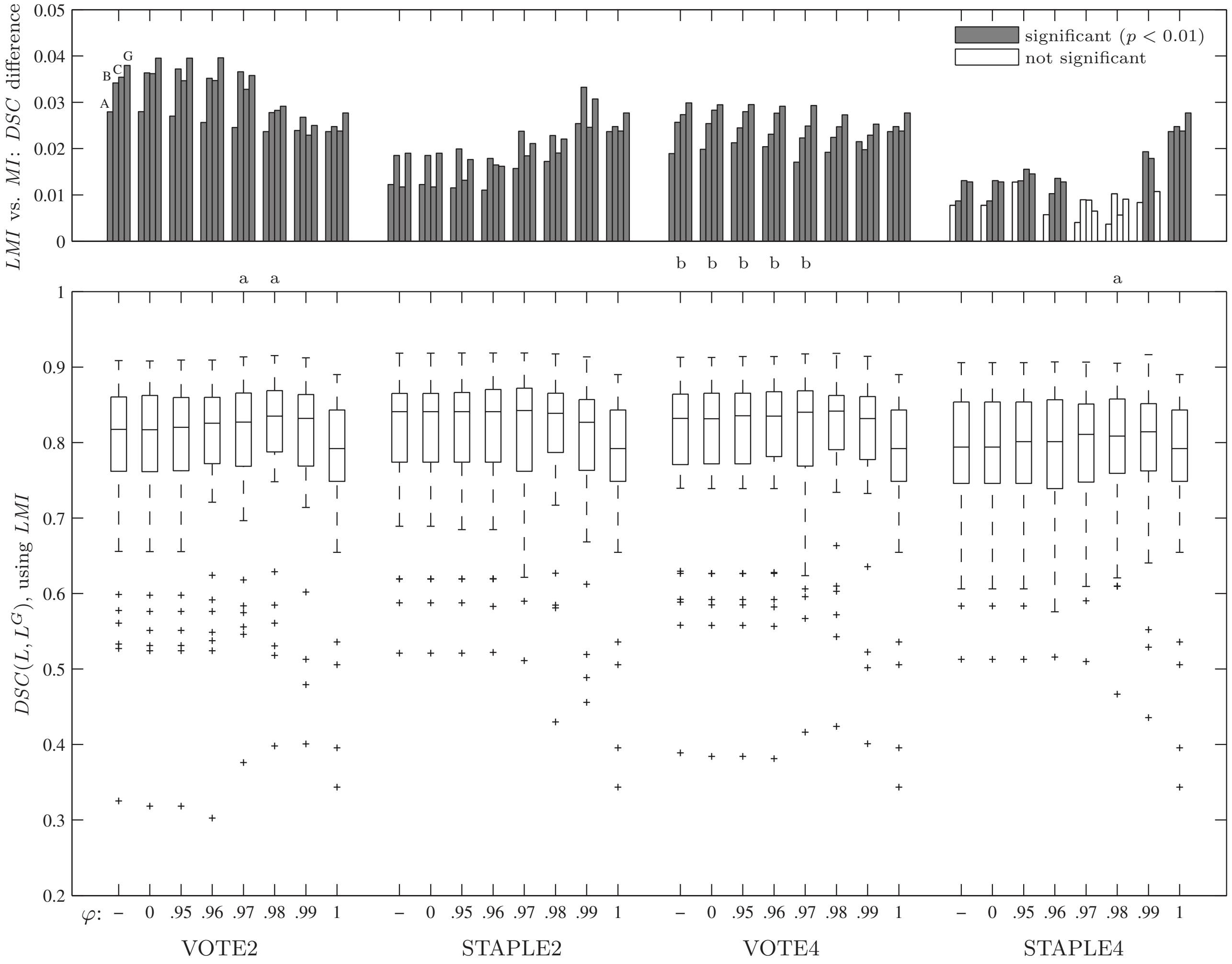


FIG. 4. Results for Experiment I. The lower part of the figure shows the effect of the four label image fusion methods, each for a range of values of φ, using LMI as the similarity measure. The upper part of the graph visualizes the difference between using LMI and MI. An "a" above the lower graph indicates significant ($p<0.05$) improvement compared to φ= "-" with the same label fusion method. A "b" indicates significant ($p<0.05$) improvement compared to VOTE2 with the same value of φ.

III.C. Evaluation measures

The results are evaluated by comparing the automatically generated prostate segmentations with the manual segmentations. A well-known measure of segmentation overlap is the Dice similarity coefficient (DSC)[19]

$$DSC(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{8}$$

where X and Y represent binary label images, and $|\cdot|$ denotes the number of voxels that equal 1. A higher DSC indicates a better correspondence. A value of 1 indicates perfect overlap, a value of 0 means no overlap at all.

The DSC does not provide insight into the spatial distribution of the segmentation errors. To visualize the segmentation accuracy we use a spherical coordinate mapping of the prostate boundary.[3,7] The shortest Euclidean distance between the manual and automatic segmentation boundaries is computed for every point on the boundary of the manual segmentation. A cartographic "Mollweide equal area" projection is used to display the result, as proposed in Ref. 7.

IV. RESULTS

The experiments were performed using MI and LMI, with VOTE2, VOTE4, STAPLE2, and STAPLE4, and with different thresholds for atlas selection (φ). The DSC values between the automatic segmentation L and the expert segmentations L^A, L^B, L^C, and L^G were computed for all 50 test scans. Figures 4 and 5 summarize the results of Experiment I and Experiment II, respectively. Each box and whisker in the lower parts of the figures visualizes the distribution of DSC(L,L^G) for a specific value of φ, when using LMI for the registration. The φ values can be found on the horizontal axis. The "-" symbol refers to φ=0 combined with w_i=1, i.e., no atlas selection and equal weights. In all other cases w_i=r_i was used. Values of φ between 0 and 0.95 are not shown, since φ=0.95 was already almost equivalent to φ=0. An "a" above the lower graph indicates significant im-

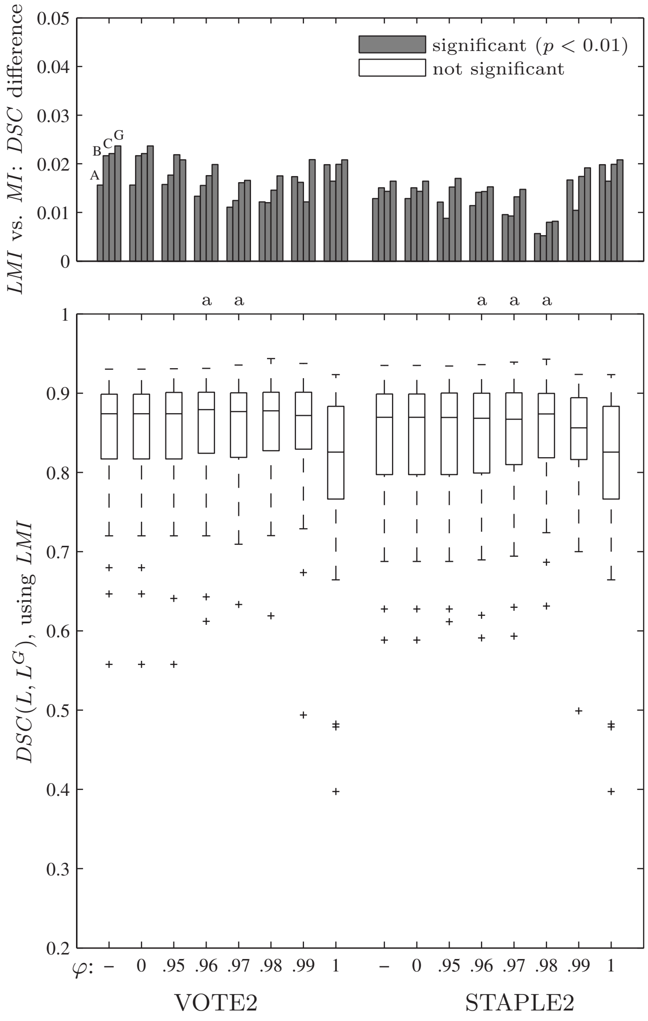


FIG. 5. Results for Experiment II. The lower part of the figure shows the effect of the two label image fusion methods, each for a range of values of φ, using LMI as the similarity measure. The upper part of the graph visualizes the difference between using LMI and MI. An "a" above the lower graph indicates significant ($p<0.05$) improvement compared to φ= "-" with the same label fusion method. A "b" would indicate significant ($p<0.05$) improvement compared to VOTE2 with the same value of φ, but this situation never occurred.

provement ($p<0.05$, a value $p<0.01$ never occurred) compared to φ="-" with the same label fusion method. A "b" indicates significant improvement ($p<0.05$) compared to VOTE2 with the same value of φ. Statistical significance was evaluated using a paired two-sided Wilcoxon test. The upper parts of Figs. 4 and 5 show the effect of LMI compared to MI. Each group of four bars displays the medians of the differences $\mathrm{DSC}(L^{LMI},L^E)-\mathrm{DSC}(L^{MI},L^E)$, for $L^E \in \{L^A,L^B,L^C,L^G\}$. Gray bars indicate that the difference is significant according to a paired two-sided Wilcoxon statistical test ($p<0.01$).

The upper parts of the figures clearly show that LMI outperformed MI in this application. The median DSC difference was positive (favoring LMI) for all settings of φ, with all tested label image fusion methods, both in Experiment I and Experiment II. Also, the choice of ground truth (L^A, L^B, L^C, or L^G) did not change the conclusion. Almost all differences were significant with $p<0.01$. Only in combination with STAPLE4 the difference was not always significant. However, the lower graph shows that STAPLE4 produced the worst results of all label image fusion methods in Experiment I. The advantage of LMI comes at no additional computational costs, thanks to the stochastic optimization method, as explained in Sec. II A. The measured computation time was around 15 min per registration on a single processor Pentium 2.8 GHz personal computer. For the implementation of LMI in Ref. 13 a computation time of 1–2 h per registration is reported. In Ref. 12 the authors report a computation time of 30 min on a cluster of 24 processors.

The lower graphs in Figs. 4 and 5 show that the differences between the different label image fusion methods were mostly rather small, but statistically significant in some cases. Selecting only the most similar atlas ($\varphi=1$) gave the worst results, which confirms the results found in Ref. 8. The optimal value of φ was around 0.98, for both the VOTE and STAPLE methods. With this value, on average 22 out of 38 atlas images were selected in Experiment I and 23 out of 49 in Experiment II. In contrast to the results reported in Ref. 18, the STAPLE algorithm did not clearly improve upon the VOTE method in our application. In Experiment I, STAPLE2 yielded somewhat better results than VOTE2 for $\varphi<0.96$, but the difference was not statistically significant. For higher values of φ, STAPLE2 and VOTE2 performed equally. STAPLE4 gave worse results than VOTE4 for all φ. In Experiment II, for all values of φ, STAPLE2 performed slightly worse than VOTE2. The additionally labeled structures in the atlas set (rectum and bladder) taken into account by VOTE4 and STAPLE4 did not lead to consistently better results either. VOTE4 improved slightly upon VOTE2, but the difference was significant only for $\varphi<0.98$. STAPLE4 performed worse than STAPLE2 for all φ. For further evaluation we use VOTE2 with $\varphi=0.98$ and LMI as the similarity measure.

Figure 6 compares the automatic segmentation results with the interobserver variability. Each box and whisker visualizes the distribution of DSC values over all 50 patients. For Experiments I and II the distributions of $\mathrm{DSC}(L,L^E)$ for $L^E \in \{L^A,L^B,L^C,L^G\}$ are shown. In Experiment I, the median DSC varied between 0.80 (with L^C) and 0.84 (with L^G). In Experiment II, the median varied between 0.85 (with L^A, L^B, and L^C) and 0.88 (with L^G). When comparing the results of Experiment I and Experiment II the very large impact of the atlas composition becomes clear. The median difference between the corresponding DSC values of Experiment I and II were all significant with $p<0.001$, according to a paired two-sided Wilcoxon statistical test. The interobserver overlap values $\mathrm{DSC}(L^A,L^B)$, $\mathrm{DSC}(L^A,L^C)$, and $\mathrm{DSC}(L^B,L^C)$ had median values of around 0.87. The results of Experiment II thus approached the level of the interobserver variability, although the human observers remained superior. The overlap of the expert segmentations with L^G was highest, which is not surprising, since L^G was constructed from L^A, L^B, and L^C

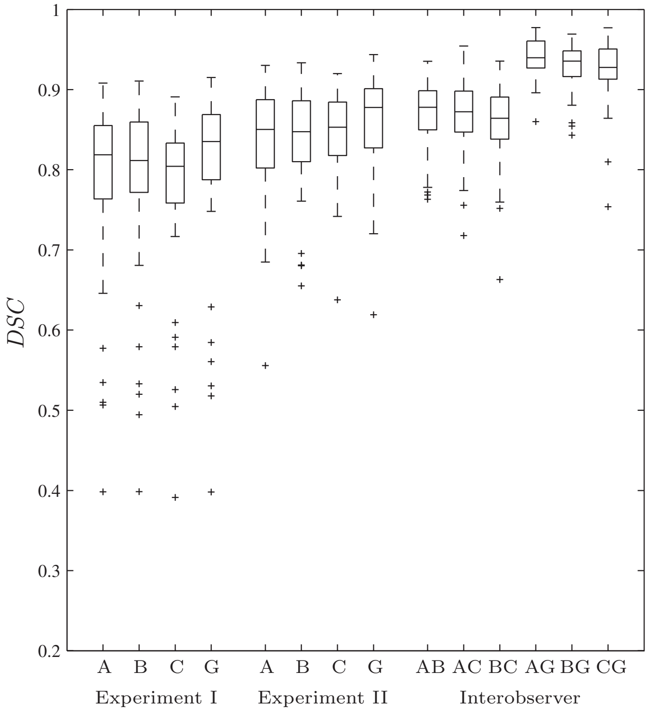


FIG. 6. The results of Experiments I and II compared to the interobserver variability. The automatic segmentation results shown were generated using LMI, VOTE2, and φ=0.98.

by majority voting. Among the three experts, observer A had the highest median DSC with L^G. Based on this and on the fact that observer A was the most experienced one, we choose to use L^A as a ground truth in the following analysis of the spatial distribution of segmentation errors.

Figure 8 shows the spatial distribution of the segmentation errors. A Mollweide map of the prostate surface is given in Fig. 7. For each test scan, the shortest Euclidean distance between the boundaries of L and L^A was computed at every point on the boundary of L^A. Subsequently, the computed distances were projected on the Mollweide map. The results of the 50 test scans were summarized by computing the first quartile, median, and third quartile of the distance at every location on the Mollweide map. Figures 8(a)–8(c) and 8(d)–8(f) show the results for Experiment I and Experiment II, respectively. In order to assess the interobserver variation, the distances between L^B and L^A and between L^C and L^A were also calculated. These results are shown in Figs. 8(g)–8(l). Note that different color scales are used for the first quartile, median, and third quartile plots.

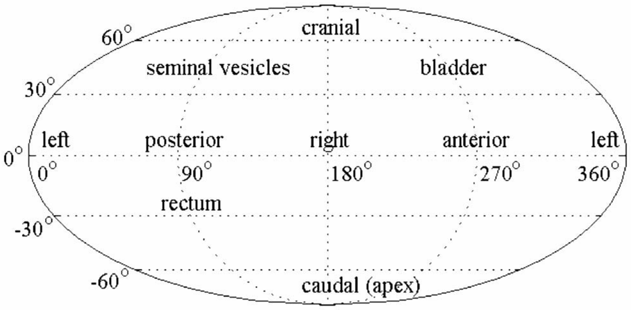


FIG. 7. Mollweide projection of the prostate boundary.

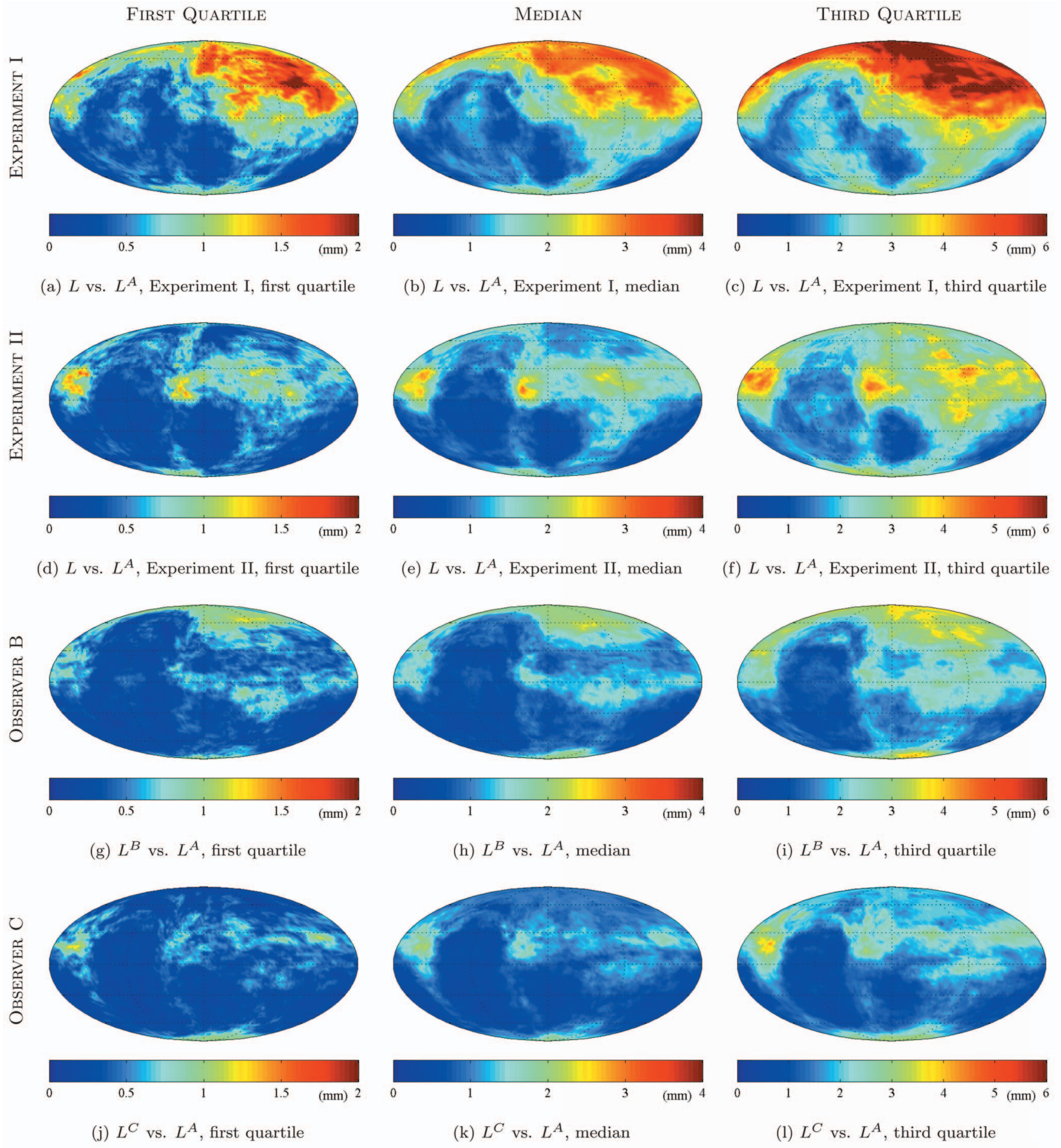
From the figures it is evident that in Experiment I the largest errors occurred at the border between the prostate and the bladder. In Experiment II the errors at the prostate-bladder boundary were much smaller, and were even comparable to the interobserver distance between L^B and L^A. The most serious segmentation errors in Experiment II were made in the tips of the seminal vesicles, which was confirmed by visual inspection of the segmentations. Both for the automatic segmentations and for the experts, the errors at the apex were relatively high. Somewhat larger errors were also observed at the anterior side of the prostate. At the prostate-rectum interface observer B and C were very close to observer A; Figures 8(i) and 8(l) show that in 75% of the cases the deviation remained below 1 mm. In Experiment II, the automatic segmentation errors at the prostate-rectum interface remained below 1 mm in 50% of the cases and below 1.5 mm in 75% of the cases, as shown in Figs. 8(e) and 8(f), respectively.

V. DISCUSSION

The accuracy of the automatic segmentation method is on a large part of the prostate surface close to the level of interobserver variability, for most test images. Most serious errors occurred around the tips of the seminal vesicles and at the anterior side of the prostate. The automatic method showed especially good performance at the prostate-rectum interface, although the human observers remained superior in most cases. Whereas a segmentation error of a few millimeters is clinically acceptable at boundaries with muscular tissue, the interfaces with rectum and bladder need to be delineated with an accuracy equal to the level of interobserver variability. Further improvement of the method is, thus, necessary.

Visual inspection of the segmentations revealed that the large errors at the prostate-bladder boundary in Experiment I (see Fig. 8) mainly occurred when the patient's prostate was very large. The volunteer data set, which is used as an atlas in Experiment I, does not contain any examples of large prostates, as shown in Fig. 3. Matching the atlas images to the patient image is thus likely to fail. A large number of the outliers observed in Fig. 6 for Experiment I can be attributed to this. The large differences between the results of Experiments I and II emphasize the importance of a proper atlas composition. Therefore, we expect that the automatic segmentation results can be further improved by explicit optimization of the atlas composition in an initial training procedure.

In previous work,[20] we have investigated to put more weight on the prostate region during registration, as a possible way to improve the results, by defining a narrow region of interest around the prostate segmentation in the atlas scan, and use only that part for registration. Experiments on the volunteer data set (used also in the current manuscript)



(a) L vs. L^A, Experiment I, first quartile

(b) L vs. L^A, Experiment I, median

(c) L vs. L^A, Experiment I, third quartile

(d) L vs. L^A, Experiment II, first quartile

(e) L vs. L^A, Experiment II, median

(f) L vs. L^A, Experiment II, third quartile

(g) L^B vs. L^A, first quartile

(h) L^B vs. L^A, median

(i) L^B vs. L^A, third quartile

(j) L^C vs. L^A, first quartile

(k) L^C vs. L^A, median

(l) L^C vs. L^A, third quartile

FIG. 8. The spatial distribution of automatic segmentation errors and interobserver variation. Figure 7 shows a map of the prostate surface that aids in interpretation. Note that the graphs have different color scales. The automatic segmentation results shown were generated using LMI, VOTE2, and $\varphi=0.98$.

showed some improvement over standard MI registration, but the use of LMI showed superior results on the same data set.[21]

The MR images used to evaluate our algorithm were acquired using a bSSFP sequence with fat suppression. Using this sequence, a high resolution ($0.49\times0.49\times1.0$ mm) is obtained in a scan time of about 2 min on a 3T MR scanner. The protocol was optimized for maximum contrast, to facilitate manual prostate contouring. It remains to be investigated whether the protocol is optimal for *automatic* prostate segmentation. The bSSFP sequence is sensitive to susceptibility artifacts, for instance, due to the presence of air in the rectum. These artifacts disturbed the nonrigid registration procedure in some cases.

The clinical test images from the patient data set are planning scans of patients scheduled for *external beam* radiation therapy. In clinical practice, the calculations of the dose distributions are actually based on these scans. In the case of *brachytherapy*, the dose plan may be adapted based on intraoperative scans. We expect that the deformations induced by the insertion of seeds do not form a problem, given the already large anatomical variability that the method is able to cope with, and the varying states of rectum and bladder filling. Susceptibility artifacts in the images due to the presence of the seeds will become an additional challenge though.

To the best of our knowledge, no other automatic segmentation results for the prostate including the seminal vesicles on 3D MR scans are available in the literature. In Ref. 4 a semiautomatic segmentation method is presented for the prostate without seminal vesicles. The method is evaluated on 3D MR scans of 24 patients. The results are given in terms of the "volume overlap" between manual and automatic segmentations. The volume overlap is also known as the Tannimoto coefficient (TC) and is related to the DSC by $\mathrm{DSC}=2\mathrm{TC}/(\mathrm{TC}+1)$.[22] A mean TC of 0.78 is reported with standard deviation ±0.05, which corresponds to a DSC of 0.88 ± 0.04. This is somewhat better than our results in Fig. 6 for Experiment II. However, the presence of the seminal vesicles increases the surface-to-volume ratio of the segmented structure, which increases the sensitivity of the DSC measure.[8] In Ref. 6, a pseudo-3D active shape modeling approach is used to segment the prostate without seminal vesicles. The method is validated on 26 3D MR scans, on a slice-by-slice basis, using the root mean square distance (RMSD) between the manual and automatic segmentation. A mean RMSD of 5.5 mm with a standard deviation of ±2.9 mm is reported. We may compare this result to Fig. 8(f), which shows that, with our method, the segmentation error remained at every location below 5 mm in 75% of the test cases. While this result seems to be in favor of our method, the RMSD values reported in Ref. 6 might be lower if they would not have been computed on a slice-by-slice basis, but in 3D. Also, it should be noted that both methods mentioned above[4,6] were validated on MR scans, acquired using a 1.5 T machine, with highly anisotropic voxels.

VI. CONCLUSION

An automatic prostate segmentation method for pelvic MR images has been proposed. The method is based on matching of manually segmented atlas images. To account for the large variability in shape, multiple atlas images are combined. A computationally efficient localized mutual information similarity measure is used in the matching stage. Evaluation was performed on a set of 50 clinical scans, which were manually segmented by three experts.

The choice of similarity measure and the composition of the atlas were demonstrated to be important determinants of segmentation quality. Using localized mutual information instead of standard mutual information yielded a significant ($p<0.01$) improvement of around 0.02, in terms of the median Dice similarity coefficient (DSC). Using an atlas composed of patient data instead of volunteer data resulted in a median DSC increase of 0.04 (significant with $p<0.01$), accompanied by a great reduction of the number of outliers. The label image fusion procedure had only a modest influence on the results. A majority voting method with an atlas selection level of $\varphi=0.98$ gave good results.

With the best settings, a median DSC of around 0.85 was achieved for the prostate, which is close to the interobserver variability of 0.87. The segmentation quality was especially good at the prostate-rectum interface, where the segmentation error remained below 1 mm in 50% of the cases and below 1.5 mm in 75% of the cases.

ACKNOWLEDGMENTS

The authors thank Ellen Kerkhof for providing the segmentations of the volunteer data. Funding for this research has been provided by the Netherlands Organisation for Scientific Research (NWO). Additionally, this work benefited from the use of the Insight Segmentation and Registration Toolkit (ITK), an open source software developed as an initiative of the U.S. National Library of Medicine and available at www.itk.org. The Mollweide projections (Fig. 8) were generated using the "M_Map" package, by Rich Pawlowicz, available at www.eos.ubc.ca/~rich.

[a] Author to whom correspondence should be addressed. Electronic mail: stefan@isi.uu.nl

[1] G. M. Villeirs, K. Van Vaerenbergh, L. Vakaet, S. Bral, F. Claus, W. J. De Neve, K. L. Verstraete, and G. O. De Meerleer, "Interobserver delineation variation using CT versus combined CT+MRI in intensity-modulated radiotherapy for prostate cancer," Strahlenther. Onkol. **181**(7), 424–430 (2005).

[2] G. M. Villeirs and G. O. De Meerleer, "Magnetic resonance imaging (MRI) anatomy of the prostate and application of MRI in radiotherapy planning," Eur. J. Radiol. **63**(3), 361–368 (2007).

[3] C. Rasch, I. Barillot, P. Remeijer, A. Touw, M. van Herk, and J. V. Lebesque, "Definition of the prostate in CT and MRI: A multi-observer study," Int. J. Radiat. Oncol., Biol., Phys. **43**(1), 57–66 (1999).

[4] D. Pasquier, T. Lacornerie, M. Vermandel, J. Rousseau, E. Lartigau, and N. Bertrouni, "Automatic segmentation of pelvic structures from magnetic resonance images for prostate cancer radiotherapy," Int. J. Radiat. Oncol., Biol., Phys. **68**(2), 592–600 (2007).

[5] Y. Zhu, S. Williams, and R. Zwiggelaar, "Computer technology in detection and staging of prostate carcinoma: A review," Med. Image Anal. **10**, 178–199 (2006).

[6] Y. Zhu, S. Williams, and R. Zwiggelaar, "A hybrid ASM approach for sparse volumetric data segmentation," Pattern Recogn. Image Anal. **17**(2), 252–258 (2007).

[7] M. Foskey, B. Davis, L. Goyal, S. Chang, E. Chaney, N. Strehl, S. Tomei, J. Rosenman, and S. Joshi, "Large deformation three-dimensional image registration in image-guided radiation therapy," Phys. Med. Biol. **50**, 5869–5892 (2005).

[8] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, Jr., "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," NeuroImage **21**(4), 1428–1442 (2004).

[9] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," IEEE Trans. Med. Imaging **18**(8), 712–721 (1999).

[10] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," IEEE Trans. Med. Imaging **16**(2), 187–198 (1997).

[11] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," Int. J. Comput. Vis. **24**(2), 137–154 (1997).

[12] G. Hermosillo, Variational methods for multimodal image matching,

Ph.D. thesis, Université de Nice, Sophia Antipolis, (2002).

[13] C. Studholme, C. Drapaca, B. Iordanova, and V. Cardenas, "Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change," IEEE Trans. Med. Imaging **25**(5), 626–639 (2006).

[14] S. Klein, M. Staring, and J. P. W. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines," IEEE Trans. Image Process. **16**(12), 2879–2890 (2007).

[15] P. Thévenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," IEEE Trans. Image Process. **9**(12), 2083–2099 (2000).

[16] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignments," Pattern Recogn. **32**, 71–86 (1999).

[17] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," IEEE Trans. Med. Imaging **23**(7), 903–921 (2004).

[18] T. Rohlfing, D. B. Russakoff, and C. R. Maurer, Jr., "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation," IEEE Trans. Med. Imaging **23**(8), 983–994 (2004).

[19] L. R. Dice, "Measures of the amount of ecologic association between species," Ecology **26**(3), 297–302 (1945).

[20] S. Klein, U. A. van der Heide, B. W. Raaymakers, A. N. T. J. Kotte, M. Staring, and J. P. W. Pluim, "Segmentation of the prostate in MR images by atlas matching," In J. A. Fessler and T. S. Denney, Jr., editors, *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 1300–1303 (2007).

[21] S. Klein, U. A. van der Heide, M. Staring, A. N. T. J. Kotte, B. W. Raaymakers, and J. P. W. Pluim, "Segmentation of the prostate in MR images by atlas matching using localized mutual information," In D. A. Jaffray, M. Sharpe, J. van Dyk, and J. P. Bissonnette, editors, *XVth International Conference on the Use of Computers in Radiation Therapy*, **2**, 585–589 (2007).

[22] W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," IEEE Trans. Med. Imaging **25**(11) 1451–1461 (2006).