Medical Image Analysis 18 (2014) 359–373

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge

Geert Litjens [a,*], Robert Toth [b], Wendy van de Ven [a], Caroline Hoeks [a], Sjoerd Kerkstra [a], Bram van Ginneken [a], Graham Vincent [e], Gwenael Guillard [e], Neil Birbeck [f], Jindang Zhang [f], Robin Strand [g], Filip Malmberg [g], Yangming Ou [h], Christos Davatzikos [h], Matthias Kirschner [i], Florian Jung [i], Jing Yuan [j], Wu Qiu [j], Qinquan Gao [k], Philip "Eddie" Edwards [k], Bianca Maan [l], Ferdinand van der Heijden [l], Soumya Ghose [d,m,n], Jhimli Mitra [d,m,n], Jason Dowling [d], Dean Barratt [c], Henkjan Huisman [a], Anant Madabhushi [b]

[a] Radboud University Nijmegen Medical Centre, The Netherlands
[b] Case Western Reserve University, USA
[c] University College London, England, United Kingdom
[d] Commonwealth Scientific and Industrial Research Organisation, Australia
[e] Imorphics, England, United Kingdom
[f] Siemens Corporate Research, USA
[g] Uppsala University, Sweden
[h] University of Pennsylvania, USA
[i] Technische Universitat Darmstadt, Germany
[j] Robarts Research Institute, Canada
[k] Imperial College London, England, United Kingdom
[l] University of Twente, The Netherlands
[m] Université de Bourgogne, France
[n] Universitat de Girona, Spain

ARTICLE INFO

Article history:
Received 17 April 2013
Received in revised form 3 December 2013
Accepted 5 December 2013
Available online 25 December 2013

Keywords:
Segmentation
Prostate
MRI
Challenge

ABSTRACT

Prostate MRI image segmentation has been an area of intense research due to the increased use of MRI as a modality for the clinical workup of prostate cancer. Segmentation is useful for various tasks, e.g. to accurately localize prostate boundaries for radiotherapy or to initialize multi-modal registration algorithms. In the past, it has been difficult for research groups to evaluate prostate segmentation algorithms on multi-center, multi-vendor and multi-protocol data. Especially because we are dealing with MR images, image appearance, resolution and the presence of artifacts are affected by differences in scanners and/or protocols, which in turn can have a large influence on algorithm accuracy. The Prostate MR Image Segmentation (PROMISE12) challenge was setup to allow a fair and meaningful comparison of segmentation methods on the basis of performance and robustness. In this work we will discuss the initial results of the online PROMISE12 challenge, and the results obtained in the live challenge workshop hosted by the MICCAI2012 conference. In the challenge, 100 prostate MR cases from 4 different centers were included, with differences in scanner manufacturer, field strength and protocol. A total of 11 teams from academic research groups and industry participated. Algorithms showed a wide variety in methods and implementation, including active appearance models, atlas registration and level sets. Evaluation was performed using boundary and volume based metrics which were combined into a single score relating the metrics to human expert performance. The winners of the challenge where the algorithms by teams Imorphics and ScrAutoProstate, with scores of 85.72 and 84.29 overall. Both algorithms where significantly better than all other algorithms in the challenge ($p < 0.05$) and had an efficient implementation with a run time of 8 min and 3 s per case respectively. Overall, active appearance model based approaches seemed to outperform other approaches like multi-atlas registration, both on accuracy and computation time. Although

* Corresponding author. Tel.: + 31243655793.
E-mail address: g.litjens@rad.umcn.nl (G. Litjens).
URL: http://www.diagnijmegen.nl (G. Litjens).

1361-8415/$ - see front matter
http://dx.doi.org/10.1016/j.media.2013.12.002

average algorithm performance was good to excellent and the Imorphics algorithm outperformed the second observer on average, we showed that algorithm combination might lead to further improvement, indicating that optimal performance for prostate segmentation is not yet obtained. All results are available online at http://promise12.grand-challenge.org/.

1. Introduction

Prostate MRI image segmentation has been an area of intense research due to the increased use of MRI as a modality for the clinical workup of prostate cancer, e.g. diagnosis and treatment planning (Tanimoto et al., 2007; Kitajima et al., 2010; Villeirs et al., 2011; Hambrock et al., 2012; Hoeks et al., 2013). Segmentation is useful for various tasks: to accurately localize prostate boundaries for radiotherapy (Pasquier et al., 2007), perform volume estimation to track disease progression (Toth et al., 2011a), to initialize multimodal registration algorithms (Hu et al., 2012) or to obtain the region of interest for computer-aided detection of prostate cancer (Vos et al., 2012; Tiwari et al., 2013), among others. As manual delineation of the prostate boundaries is time consuming and subject to inter- and intra-observer variation, several groups have researched (semi-) automatic methods for prostate segmentation (Pasquier et al., 2007; Costa et al., 2007; Klein et al., 2008; Makni et al., 2009; Toth et al., 2011b; Chandra et al., 2012; Gao et al., 2012b). However, as most algorithms are evaluated on proprietary datasets a meaningful comparison is difficult to make.

This problem is aggravated by the fact that most papers cannot include a comparison against the state-of-the-art due to previous algorithms being either closed source or very difficult to implement without help of the original author. Especially in MRI, where signal intensity is not standardized and image appearance is for a large part determined by acquisition protocol, field strength, coil profile and scanner type, these issues present a major obstacle in further development and improvement of prostate segmentation algorithms.

In recent years several successful 'Grand Challenges in Medical Imaging' have been organized to solve similar issues in the fields of liver segmentation on CT (Heimann et al., 2009), coronary image analysis (Schaap et al., 2009), brain segmentation on MR (Shattuck et al., 2009), retinal image analysis (Niemeijer et al., 2010) and lung registration on CT (Murphy et al., 2011). The general design of these challenges is that a large set of representative training data is publicly released, including a reference standard for the task at hand (e.g. liver segmentations). A second set is released to the public without a reference standard, the test data. The reference standard for the test data is used by the challenge organizers to evaluate the algorithms. Contestants are then allowed to tune their algorithms to the training data after which their results on the test data are submitted to the organizers who calculate predefined evaluation measures on these test results. The objective of most challenges is to provide independent evaluation criteria and subsequently rank the algorithms based on these criteria. This approach overcomes the usual disadvantages of algorithm comparison, in particular, bias.

The Prostate MR Image Segmentation (PROMISE12) challenge presented in this paper tries to standardize evaluation and objectively compare algorithm performance for the segmentation of prostate MR images. To achieve this goal a large, representative set of 100 MR images was made available through the challenge website: http://promise12.grand-challenge.org/. This set was subdivided into training (50), test (30) and live challenge (20) datasets (for further details on the data, see Section 2). Participants could download the data and apply their own algorithms. The goal of the challenge was to accurately segment the prostate capsule. The calculated segmentations on the test set were then submitted to the challenge organizers through the website for independent evaluation. Evaluation of the results included both boundary and volume based metrics to allow a rigorous assessment of segmentation accuracy. To calculate an algorithm score based on these metrics, they were compared against human readers. Further details about generation of the algorithm score can be found in Section 3.2.

This paper will describe the setup of the challenge and the initial results obtained prior to and at the workshop hosted by the MICCAI2012 conference in Nice, where a live challenge was held between all participants. New results, which can still be submitted through the PROMISE12 website, can be viewed online.

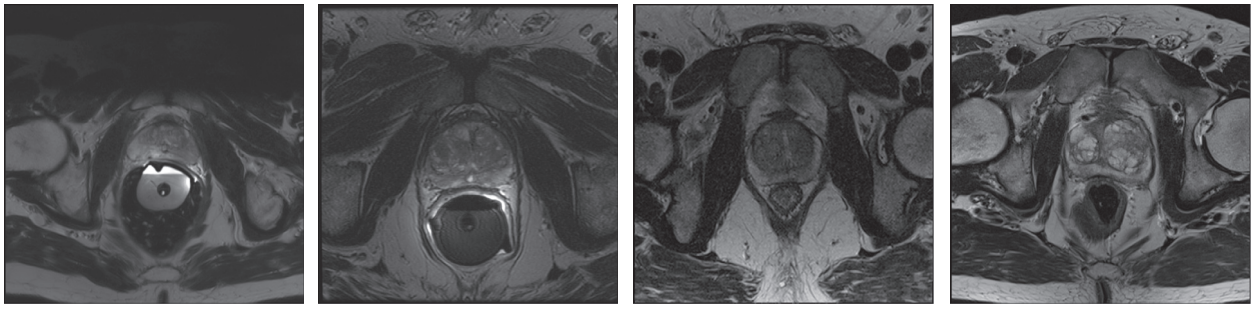
2. Materials

2.1. MRI images

In MRI images, the pixel/voxel intensities and therefore appearance characteristics of the prostate can greatly differ between acquisition protocols, field strengths and scanners (Dickinson et al., 2011; Barentsz et al., 2012). Example causes of appearance differences include the bias field (Leemput et al., 1999; Sung et al., 2013), signal-to-noise ratio (Fütterer and Barentsz, 2009; Li et al., 2012) and resolution (Tanimoto et al., 2007; Kitajima et al., 2010), especially through-plane. Additionally, signal intensity values are not standardized (Nyúl and Udupa, 1999; Nyúl et al., 2000). Therefore a segmentation algorithm designed for use in clinical practice needs to deal with these issues (Bezdek et al., 1993; Clarke et al., 1993). Consequently, we decided to include data from four different centers: Haukeland University Hospital (HK) in Norway, the Beth Israel Deaconess Medical Center (BIDMC) in the US, University College London (UCL) in the United Kingdom and the Radboud University Nijmegen Medical Centre (RUNMC) in the Netherlands. Each of the centers provided 25 transverse T2-weighted MR images. This resulted in a total of 100 MR images. Details pertaining to the acquisition can be found in Table 1. Additionally, a central slice of a data set for each of the centers is shown in Fig. 1 to show the appearance differences. These scans where acquired either for prostate cancer detection or staging purposes. However, the clinical stage of the patients and the presence and location of prostate cancer is unknown to the organizers. Transverse T2-weighted MR was used because these contain most anatomical detail (Barentsz et al., 2012), are used clinically for prostate volume measurements (Hoeks et al., 2013; Toth et al., 2011a) and because most current research papers focus on

Table 1
Details of the acquisition protocols for the different centers. Each center supplied 25 T2-weighted MR images of the prostate.

Center	Field strength (T)	Endorectal coil	Resolution (in-plane/through-plane in mm)	Manufacturer
HK	1.5	Yes	0.625/3.6	Siemens
BIDMC	3	Yes	0.25/2.2–3	GE
UCL	1.5 and 3	No	0.325–0.625/3–3.6	Siemens
RUNMC	3	No	0.5–0.75/3.6–4.0	Siemens



(a) Haukeland University Hospital, Norway: 1.5T with endorectal coil

(b) Beth Israel Deaconess Medical Center. USA: 3.0T with endorectal coil

(c) University College London, United Kingdom: 1.5T and 3.0T without endorectal coil

(d) Radboud University Medical Centre, The Netherlands: 3.0T without endorectal coil

Fig. 1. Slice of a data set from different centers to show appearance differences.

segmentation on T2-weighted MRI. The data were then split randomly into 50 training cases, 30 test cases and 20 live challenge cases. Although the selection process was random, it was stratified according to the different centers to make sure no training bias towards a certain center could occur.

2.2. Segmentation reference standard

Each center provided a reference segmentation of the prostate capsule performed by an experienced reader. All annotations were performed on a slice-by-slice basis using a contouring tool. The contouring tool itself was different for the different institutions, but the way cases were contoured was similar. Contouring was performed by annotating spline-connected points in either 3DSlicer (www.slicer.org) or MeVisLab (www.mevislab.de). The reference segmentations were checked by a second expert, C.H., who has read more than 1000 prostate MRIs, to make sure they were consistent. This expert had no part in the initial segmentation of the cases and was asked to correct the segmentation if inconsistencies were found. The resulting corrected segmentations were used as the reference standard segmentation for the challenge. An example of a reference segmentation at the base, center and apex of the prostate is shown in Fig. 2.

2.3. Second observer

For both the testing and the live challenge data a relatively inexperienced nonclinical observer (W.v.d.V, two years of experience with prostate MR research) was asked to manually segment the prostate capsule using a contouring tool. The second observer was blinded to the reference standard to make sure both segmentations were independent. The second observer segmentations were used to transform the evaluation metrics into a case score, as will be explained in Section 3.2. An example of a second observer segmentation is shown in Fig. 2.

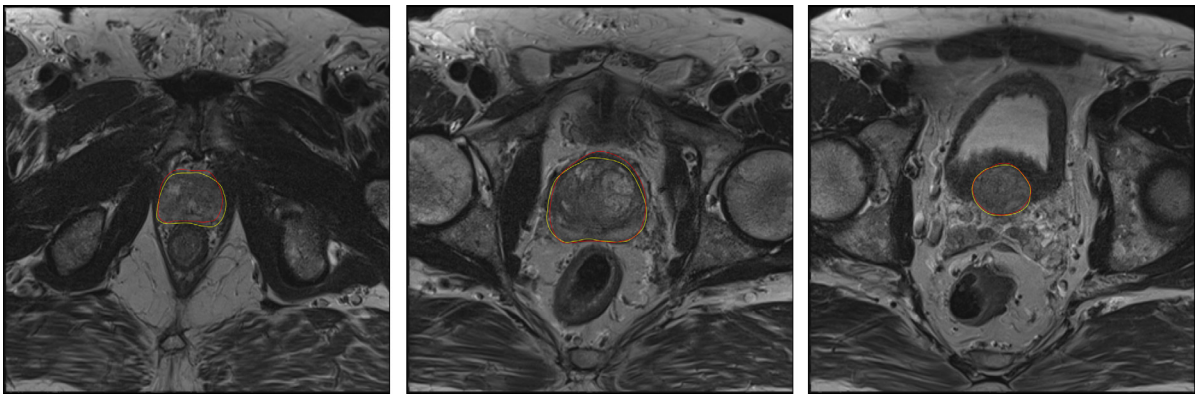
3. Evaluation

3.1. Metrics

The metrics used in this study are widely used for the evaluation of segmentation algorithms:

1. the Dice coefficient (DSC) (Klein et al., 2008; Heimann et al., 2009),
2. the absolute relative volume difference, the percentage of the absolute difference between the volumes (aRVD) (Heimann et al., 2009),
3. the average boundary distance, the average over the shortest distances between the boundary points of the volumes (ABD) (Heimann et al., 2009),
4. the 95% Hausdorff distance (95HD) (Chandra et al., 2012).

All evaluation metrics were calculated in 3D. We chose both boundary and volume metrics to give a more complete view of segmentation accuracy, i.e. in radiotherapy boundary based metrics



(a) Apex (b) Center (c) Base

Fig. 2. Example T2-weighted transverse prostate MRI images displaying an apical, central and basal slice. The reference standard segmentation is shown in yellow and the second observer segmentation in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

would be more important, whereas in volumetry the volume metrics would be more important. In addition to evaluating these metrics over the entire prostate segmentation, we also calculated them specifically for the apex and base parts of the prostate, because these parts are very important to segment correctly, for example in radiotherapy and TRUS/MR fusion. Moreover, these are the most difficult parts to segment due the large inter-patient variability and differences in slice thickness. To determine the apex and base the prostate was divided into three approximately equal parts in the slice dimension (the caudal 1/3 of the prostate volume was considered apex, the cranial 1/3 was considered base). If a prostate had a number of slices not dividable by 3 (e.g. 14), the prostate would be divided as 4-6-4 for the base, mid-gland and apex respectively.

The DSC was calculated using:

$$D(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1}$$

where $|X|$ is the number of voxels in the reference segmentation and $|Y|$ is the number of voxels in the algorithm segmentation.

The relative volume difference was calculated as:

$$\mathrm{RVD}(\mathrm{X},\mathrm{Y}) = 100 \times \left(\frac{|X|}{|Y|} - 1\right) \tag{2}$$

and thus the absolute relative volume difference is

$$\mathrm{aRVD}(\mathrm{X},\mathrm{Y}) = |\mathrm{RVD}(X,Y)| \tag{3}$$

Note that although we use the aRVD to measure algorithm performance (both under- and over-segmentation are equally bad), in the results we will present the RVD, which makes it possible to identify if algorithms on average tend to over- or under-segment the prostate.

For both the 95th percentile Hausdorff distance and the average boundary distance we first extract the surfaces of the reference segmentation and the algorithm segmentation. The regular Hausdorff distance is then defined as:

$$\mathrm{HD}_{\mathrm{asym}}(X_s, Y_s) = \max_{x \in X_s}\left(\min_{y \in Y_s} \mathbf{d}(x,y)\right) \tag{4}$$

$$\mathrm{HD}(X_s, Y_s) = \max\left(\mathrm{HD}_{\mathrm{asym}}(X_s, Y_s), \mathrm{HD}_{\mathrm{asym}}(Y_s, X_s)\right) \tag{5}$$

where X_s and Y_s are the sets of surface points of the reference and algorithm segmentations respectively. The operator $\mathbf{d}$ is the Euclidean distance operator. As the normal Hausdorff distance is very sensitive to outliers we use the 95th percentile of the asymmetric Hausdorff distances instead of the maximum.

Finally, the average boundary distance (ABD) is defined as:

$$\mathrm{ABD}(X_s, Y_s) = \frac{1}{N_{X_s} + N_{Y_s}}\left(\sum_{x \in X_s}\min_{y \in Y_s} \mathbf{d}(x,y) + \sum_{y \in Y_s}\min_{x \in X_s} \mathbf{d}(y,x)\right) \tag{6}$$

3.2. Score

Algorithms were ranked by comparing the resulting evaluation measures to the second observer and the reference segmentation in a way similar to Heimann et al. (2009). First, the metrics of the second observer segmentations are calculated with respect to the reference segmentation. Then we average each metric over all cases and define a mapping function:

$$\mathrm{score}(x) = \max(ax + b, 0) \tag{7}$$

This function maps a metric value x to a score between 0 and 100. The equation is solved for a and b by setting a score of 100 to a perfect metric result, e.g. a DSC of 1.0 and setting a score of 85 to a metric result equal to the average metric value of the second observer. This will give us two equations to solve the two unknowns, a and b. Additionally, a score of zero was set as the minimum because otherwise cases with a very poor or missing segmentation could bias the final score of an algorithm too much. As an example, if the second observer segmentations have an average DSC of 0.83, a and b are 88.24 and 11.76 respectively. As such, if an algorithm obtains a DSC of 0.87 on a case the score will be 88.53. This approach is applied to all metrics. The scores for all metrics were averaged to obtain a score per case. Then the average over all cases was used to rank the algorithms.

A relatively high reference score of 85 was chosen for the second observer because her segmentations were in excellent correspondence with the reference standard. An even higher score than 85 would not be warranted, as the segmentations still contain errors experienced observers would not make. The average metric scores for the second observer are presented in Tables 6 and 7. Comparing these metric scores to scores reported in literature for inter-observer variability we can see that they are at approximately at the same level (Pasquier et al., 2007; Costa et al., 2007; Klein et al., 2008; Makni et al., 2009; Toth et al., 2011b; Chandra et al., 2012; Gao et al., 2012b).

The main reason to use this approach is that it allows us to incorporate completely different, but equally important metrics like average boundary distance and the Dice coefficient. Furthermore, in addition to allowing us to rank algorithms, the scores themselves are also meaningful, i.e. higher scores actually correspond to better segmentations. An alternative approach could have been to rank algorithms per metric and average the ranks over all metrics. However, such an average rank is not necessarily related to a segmentation performance: the best ranking algorithm could still show poor segmentation results that are much worse than the second observer.

4. Methods

This section gives an overview of all the segmentation methods that participated in the challenge. A short description for each algorithm is given. More detailed descriptions of the algorithms can be found in peer-reviewed papers submitted to the PROMISE12 challenge, available at: http://promise12.grand-challenge.org/Results. Algorithms were categorized as either automatic (no user interaction at all), semi-automatic (little user interaction, e.g. setting a single seed point) or interactive (much user interaction, e.g. painting large parts of the prostate). The algorithm categories and additional details can be found in Tables 2 and 8. The names in subsection titles are the team names chosen by the participants and are as such not related to the method themselves. Most names are either abbreviations of group names or company names. Links to the websites of the individual groups can also be found on the PROMISE12-website.

4.1. Fully automatic segmentation of the prostate using active appearance models – Imorphics

Vincent et al. (2012) of Imorphics Ltd. have developed a generic statistical modeling system designed to minimize any bespoke development needed for different anatomical structures and image modalities.

The Imorphics system generates a set of dense anatomical landmarks from manually segmented surfaces using a variant of the Minimum Description Length approach to Groupwise Image Registration (Cootes et al., 2005). The correspondence points and associated images are used to build an Appearance Model. The Appearance Model is matched to an unseen image using an Active Appearance Model (AAM) which optimizes the model parameters

Table 2
Overall challenge results. The last three columns contain the scores including standard deviations. These scores are an average of all individual metric scores over all cases, as explained in Section 3.2. For the live challenge scores with an asterisk, teams had either missing or incomplete segmentations for some cases. Incomplete or failed cases were assigned a score of 0. The scores of these groups over all completed cases is shown in brackets. The UBUdG team did not participate in the live challenge and as such received a zero score.

Rank	Team name	Type	Online	Live	Average
1	Imorphics	Automatic	84.36 ± 7.11	87.07 ± 3.36	85.72 ± 5.90
2	ScrAutoProstate	Automatic	83.49 ± 5.92	85.08 ± 3.54	84.29 ± 5.10
3	CBA	Interactive	80.66 ± 6.46	81.21 ± 9.60	80.94 ± 7.86
4	Robarts	Semi-automatic	77.32 ± 4.04	80.08 ± 7.18	78.70 ± 5.51
5	Utwente	Semi-automatic	75.23 ± 10.53	80.26 ± 7.30	77.75 ± 9.37
6	Grislies	Automatic	77.56 ± 12.60	74.35 ± 11.28	75.96 ± 12.08
7	ICProstateSeg	Automatic	76.06 ± 9.40	$75.74 \pm 8.81^{*}$ (84.16 ± 4.43)	75.90 ± 9.17
8	DIAG	Automatic	73.30 ± 13.69	77.01 ± 12.09	75.16 ± 13.07
9	SBIA	Automatic	78.34 ± 8.22	$61.38 \pm 28.22^{*}$ (76.72 ± 7.44)	69.86 ± 18.95
10	Rutgers	Automatic	65.97 ± 13.13	71.77 ± 11.02	68.87 ± 12.32
11	UBUdG	Semi-automatic	70.44 ± 9.12	$00.00 \pm 0.0^{*}$	35.22 ± 9.12
–	SecondObserver	–	85.00 ± 4.50	85.00 ± 4.91	85.00 ± 4.67

to generate an instance which matches the image as closely as possible (Cootes et al., 2001).

Active Appearance Models require an initial estimate of the model parameters including position, rotation and scale. The system uses a multi-resolution gridded search method. This is started at a low image and model resolution with a small number of measured residuals to make it reasonably fast. The results of these searches are ranked according to the sum of squares of the residual, and a proportion removed from consideration. The remaining search results are used to initialize models at a higher resolution, and so on. Finally, the single best result at the highest resolution gives the segmentation result.

4.2. Region-specific hierarchical segmentation of MR prostate using discriminative learning – ScrAutoProstate

The segmentation pipeline developed by Birkbeck et al. (2012) addresses the challenges of MR prostate segmentation through the use of region-specific hierarchical segmentation with discriminative learning.

First, an intensity normalization is used to adjust for global contrast changes across the images. Images with an endorectal coil are then further enhanced by flattening the intensity profile on the bright regions near the coil using an automatic application of Poisson image editing (Pérez et al., 2003).

In the next phase of the pipeline, a statistical model of mesh surface variation learned from training data is aligned to the normalized image. The pose parameters of the shape model are extracted through the use of marginal space learning (Zheng et al., 2009), which decomposes the estimation of pose into sequential estimates of the position, orientation, scale, and then the first few modes of variation. The estimation of each set of pose parameters relies on a probabilistic boosting tree classifier to discriminatively model the relationship between the image data and the unknown parameters being estimated. During training, each classifier automatically selects the most salient features from a large feature pool of Haar and steerable features. After the statistical mesh model has been aligned to the input image using marginal space learning, the segmentation is refined through a coarse-to-fine boundary refinement that uses surface varying classifiers to discriminate the boundary of the prostate from adjacent soft tissue. The mesh from this final refinement stage is constrained by the statistical shape model.

4.3. Smart paint – CBA

Malmberg et al. (2012) have developed an interactive segmentation tool called Smart Paint. The user segments the organ of interest by sweeping the mouse cursor in the object or background, similar to how an airbrush is used. Areas are painted with a semi-transparent color which gives immediate feedback in the chosen interaction plane. As the paint is applied in 3D, when the user moves to another plane using the mouse thumb-wheel the effect of the painting is seen also there.

The algorithm works by taking both the spatial distance to the cursor and the image content (intensity values) into account. The image I and the segmentation function f are mappings from elements of a three dimensional voxel set to the interval [0, 1]. A voxel x belongs to the foreground if $f(x) \geqslant 0.5$, and to the background otherwise. Initially, $f = 0$. The brush tool has a value v that is either 1 (to increase the foreground) or 0 (to increase the background). A single brush stroke centered at voxel x affects the segmentation at all nearby voxels y according to

$$f(y) \leftarrow (1 - \alpha(x,y))f(y) + \alpha(x,y)v \tag{8}$$

$$\alpha(x,y) = \beta(1 - |I(y) - I(x)|)^{k} \max\left(\frac{(r - d(x,y))}{r}, 0\right) \tag{9}$$

where $d(x,y)$ is the Euclidean distance between the voxel centers of x and y, r is the brush radius specified by the user and β and k are constants.

Additionally, the user can smooth the current segmentation using a weighted average filter. The algorithm is not very sensitive to the values selected for the β and k constants. Values for β were in the range 0.01–0.1 and for k in the range 1–5 and influence the behavior of the brush. These variables could be changed by the user.

4.4. Multi-atlas segmentation of the prostate: a zooming process with robust registration and atlas selection – SBIA

The multi-atlas based segmentation framework designed by Ou et al. (2012) automatically segments the prostate in MR images. Atlases from 50 training subjects are nonrigidly registered to the target image. The calculated deformations are used to warp expert annotated prostate segmentations of the atlases into the target image space. The warped prostate annotations are then fused by the STAPLE strategy (Warfield et al., 2004) to form a single prostate segmentation in the target image.

The main challenge in this multi-atlas segmentation framework is image registration. To account for the registration challenges, three measures are taken in the multi-atlas segmentation framework. First, the DRAMMS image registration algorithm is used (Ou et al., 2010). DRAMMS establishes anatomical correspondences by using high dimensional texture features at each voxel. Voxel texture features are more distinct than just using intensity, which

helps to improve registration accuracy. Second, a two-phase strategy is used. In phase 1 the entire prostate images from training subjects are used to compute an initial segmentation of the prostate in target image. Phase 2 focuses only on the initially segmented prostate region and its immediate neighborhood. Third, in each phase, atlas selection is used. Those atlases having high similarity with the target image in the prostate regions after registration are kept. Similarity is measured using the correlation coefficient, mutual information, as well as the DSC between the warped prostate annotation and the tentative prostate segmentation.

4.5. Automatic prostate segmentation in MR images with a probabilistic active shape model – Grislies

Kirschner et al. (2012) segment the prostate with an Active Shape Model (ASM) (Cootes et al., 2001). For training the ASM, meshes were extracted from the ground truth segmentations using Marching Cubes (Lorensen and Cline, 1987). Correspondence between the meshes was determined using a nonrigid mesh registration algorithm. The final ASM has 2000 landmarks and was trained using principal component analysis (PCA). The actual segmentation is done with a three step approach, consisting of (1) image preprocessing, (2) prostate localization and (3) adaption of the ASM to the image.

In the preprocessing step, the bias field is removed using coherent local intensity clustering, and the image intensities are normalized (Li et al., 2009). Prostate localization is done using the sliding window approach: a boosted classifier based on 3D Haar-like features is used to decide whether the subimage under the current detector window position contains the prostate or not. This approach is similar to the Viola-Jones algorithm for face detection in 2D images (Viola and Jones, 2001),.

The actual segmentation is done with a Probabilistic ASM. In this flexible ASM variant, shape constraints are imposed by minimizing an energy term which determines a compromise between three forces: an image energy that draws the model towards detected image features, a global shape energy that enforces plausibility of the shapes with respect to the learned ASM, and a local shape energy that ensures that the segmentation is smooth. For detection of the prostate's boundary, a boosted detector using 1D Haar-like features is used, which classifies sampled intensity profiles into boundary and nonboundary profiles.

4.6. An efficient convex optimization approach to 3D prostate MRI segmentation with generic star shape prior – Robarts

The work by Yuan et al. (2012) proposes a global optimization-based contour evolution approach for the segmentation of 3D prostate MRI images, which incorporates histogram matching and a variational formulation of a generic star shape prior.

The proposed method overcomes the existing challenges of segmenting 3D prostate MRIs: heterogeneous intensity distributions and a wide variety of prostate shape appearances. The proposed star shape prior does not stick to any particular object shape from learning or specified parameterized models, but potentially reduces ambiguity of prostate segmentation by ruling out inconsistent segments; it provides robustness to the segmentation when the image suffers from poor quality, noise, and artifacts.

In addition, a novel convex relaxation based method is introduced to evolve a contour to its globally optimal position during each discrete time frame, which provides a fully time implicit scheme to contour evolution and allows a large time step size to accelerate the speed of convergence.

Moreover, a new continuous max-flow formulation is proposed, which is dual to the studied convex relaxation formulation and derives a new efficient algorithm to obtain the global optimality of contour evolution. The continuous max-flow based algorithm is implemented on GPUs to significantly speed up computation in practice.

4.7. An automatic multi-atlas based prostate segmentation using local appearance specific atlases and patch-based voxel weighting – ICProstateSeg

Gao et al. (2012a) present a fully automated segmentation pipeline for multi-center and multi-vendor MRI prostate segmentation using a multi-atlas approach with local appearance specific voxel weighting.

An initial denoising and intensity inhomogeneity correction is performed on all images. Atlases are classified into two categories: normal MRI scans A_n and scans taken with a transrectal coil A_m. This is easily achieved by examining the intensity variation around the rectum since the transrectal coil produces significant physical distortion but also has a characteristic bright appearance in the local region near the coil itself. The subatlas database whose atlas appearance is closest to the new target is chosen as the initial atlas database. After that, the top N similar atlases are further chosen for atlas registration by measuring intensity difference in the region of interest around prostate.

After all the selected atlases are nonrigidly registered to a target image, the resulting transformation is used to propagate the anatomical structure labels of the atlas into the space of the target image. Finally, a patch-based local voxel weighting strategy is introduced, which was recently proposed for use in patch-based brain segmentation (Coupé et al., 2011) and improved by introducing the weight of the mapping agreement from atlas to target. After that, the label that the majority of all warped labels predict for each voxel is used for the final segmentation of the target image.

4.8. Prostate MR image segmentation using 3D active appearance models – Utwente

The segmentation method proposed by Maan and van der Heijden (2012) is an adaptation of the work presented by Kroon et al. (2012) by using a Shape Context based nonrigid surface registration in combination with 3D Active Appearance Models (AAM).

The first step in AAM training is describing the prostate surface in each training case by a set of landmarks. Every landmark in a training case must have a corresponding landmark in all other training cases. To obtain the corresponding points Shape Context based nonrigid registration of the binary segmentation surfaces was used (Maan and van der Heijden, 2012; Kroon et al., 2012). PCA is applied to determine the principal modes of the shape variation. The appearance model can be obtained in a similar way: first each training image is warped so that its points correspond to the mean shape points. Subsequently, the grey-level information of the region covered by the mean shape is sampled. After normalization, a PCA is applied to obtain the appearance model. The combined shape and appearance model can generalize to almost any valid example.

During the test phase, the AAM is optimized by minimizing the difference between the test image and the synthesized images. The mean model is initialized by manually selecting the center of the prostate based on visual inspection. Subsequently, the AAM is applied using two resolutions with both 15 iterations.

4.9. A multi-atlas approach for prostate segmentation in MR images – DIAG

Litjens et al. (2012) investigated the use of a multi-atlas segmentation method to segment the prostate using the Elastix

registration package. The method is largely based on the work of Klein et al. (2008) and Langerak et al. (2010). The 50 available training data sets are used as atlases and registered to the unseen image using localized mutual information as a metric. Localized mutual information calculates the sum of the mutual information of image patches instead of the mutual information of the entire image. This approach reduces the effect of magnetic field bias and coil profile on the image registration.

The registration process consists of two steps: first a rough initial alignment is found, after which an elastic registration is performed. The 50 registered atlases are then merged to form a signal binary segmentation using the SIMPLE optimization algorithm (Langerak et al., 2010). SIMPLE tries to automatically discard badly registered atlases in an iterative fashion using the correspondence of the atlas to the segmentation result in the previous iteration. The DSC was used as the evaluation measure in the SIMPLE algorithm.

4.10. Deformable landmark-free active appearance models: application to segmentation of multi-institutional prostate MRI data – Rutgers

Toth and Madabhushi (2012a) propose a Multi-Feature, Landmark-Free Active Appearance Model (MFA) based segmentation algorithm, based on (Toth and Madabhushi, 2012b). The MFA contains both a training module and a segmentation module. The MFA is constructed by first aligning all the training images using an affine transformation. Second, the shape is estimated by taking the signed distance to the prostate surface for each voxel, which represents a levelset, such that a value of 0 corresponds to the voxels on the prostate surface. Third, principal component analysis is used to map the shape and intensity characteristics of the set of training images to a lower dimensional space. Then a second PCA is performed on the joint set of lower dimensional shape and appearance vectors to link the shape and appearance characteristics.

To segment an unseen image, the image must be registered to the MFA, resulting in transformation T mapping the input image to the MFA. This is performed by first calculating the PCA projection of the intensities learned from the training data. Then the linked projections are reconstructed and subsequently the intensities and shape. The normalized cross-correlation between the reconstruction and the original image are calculated and the transform T is optimized to obtain maximum normalized cross-correlation. The shape corresponding to the optimal transformation was thresholded at 0 to yield the final segmentation.

While the original algorithm (Toth and Madabhushi, 2012b) defined "T" as an affine transformation, to account for the high variability in the prostate shape and appearance (e.g. with or without an endorectal coil), a deformable, b-spline based transform was used to define "T". This resulted in a more accurate registration than affine, although further studies suggest that separate subpopulation based models could potentially yield more accurate segmentations, given enough training data.

4.11. A random forest based classification approach to prostate segmentation in MRI – UBUdG

The method proposed by Ghose et al. (2012) has two major components: a probabilistic classification of the prostate and the propagation of region based levelsets to achieve a binary segmentation. The classification problem is addressed by supervised random decision forest.

During training, the number of slices in a volume containing the prostate is divided into three equal parts as apex, central and base regions. The individual slices are resized to a resolution of 256×256 pixels and a contrast-limited adaptive histogram equalization is performed to minimize the effect of magnetic field bias. Each feature vector is composed of the spatial position of a pixel and the mean and standard deviation of the gray levels of its 3×3 neighborhood. Three separate decision forests are built corresponding to the three different regions of the prostate the apex, the central region and the base. Only 50% of the available training data was used for each of the regions.

During testing the first and the last slices of the prostate are selected and the test dataset is divided into the apex, the central and the base regions. Consecutively preprocessing is done on in the same way as for the training images. Decision forests trained for each of the regions are applied to achieve a probabilistic classification of the apex, the central and the base slices. Finally evolution of the Chan and Vese levelsets on the soft classification ensures segmentation of the image into prostate and the background regions.

4.12. Combinations of algorithms

It is well known that combining the results of multiple human observers often leads to a better segmentation than using the segmentation of only a single observer (Warfield et al., 2004). To investigate whether this is also true for segmentation algorithms, different types of combinations were tried. First, combining all the algorithm results using a majority voting approach was explored. The majority voting combination considered a voxel part of the prostate segmentation if the majority of the algorithms segmented the voxel as a prostate voxel. Second, only the top 5 (expert) algorithms were combined based on the overall algorithm score. A 'best combination' reference was also included by selecting the algorithm with the maximum score per case, for both the top 5 and all algorithms.

5. Results

5.1. Online challenge

The results of the online challenge are summarized in Tables 2, 6 and Fig. 3. In Table 2 the average algorithm scores and standard deviations are presented, which are used to rank the algorithms. The ordering of the algorithms represents the ranking after both the online and live challenges. The online and live components were weighted equally to determine the final ranking. Metric values and scores for all algorithms on the online challenge data are presented in Table 6. In Fig. 3 we provide the results per algorithm per case to give a more complete view of algorithm robustness and variability.

5.2. Live challenge

Tables 2, 7 and Fig. 4 show the results of the live challenge at the MICCAI2012 workshop. In Table 2 (column 2) the average scores for each algorithm are presented including standard deviations. Metric values and scores for all algorithms on the live challenge data are presented in 7. Fig. 4 shows the scores per case per algorithm for the cases processed at the live challenge. Algorithms that were unable to segment all cases during the period of the challenge (4 h), or produced segmentations that were considered to be a failure according to algorithm-specific checking criteria or the group, are indicated with an asterisk in Table 2. Unsegmented or failed cases were given a score of 0.

5.3. Overall

The overall ranking of the algorithms is presented in Table 2. Additionally, the results of the algorithm combinations are shown

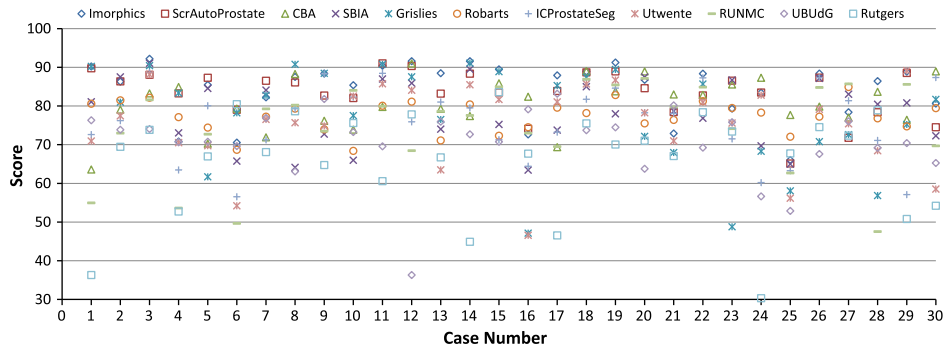


Fig. 3. Results of the online challenge. The overall score is on the vertical axis and the case number on the horizontal axis. Teams are given a different symbol and color. Case distributions per center were: 1:7 RUNMC, 8:14 BIDMC, 15:22 UCL, 23:30 HK. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

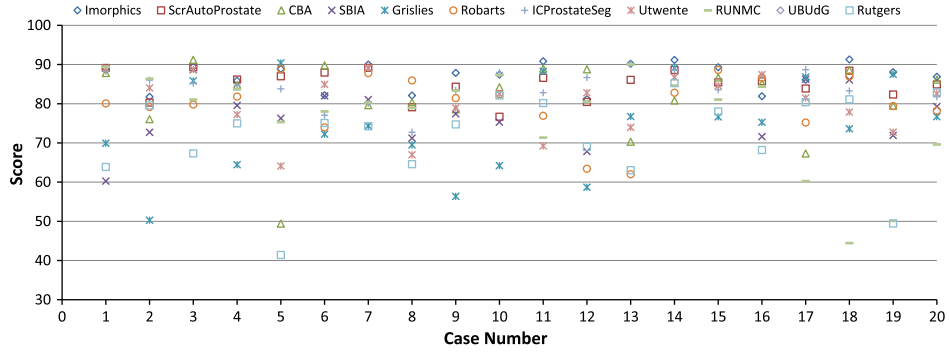


Fig. 4. Results of the live challenge. The overall score is on the vertical axis and the case number on the horizontal axis. Teams are given a different symbol and color. Case distributions per center were: 1:5 UCL, 6:10 HK, 11:15 BIDMC, 16:20 RUNMC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Results for the single best algorithm and combinations of algorithms, average over all cases including standard deviation.

Name	Online	Live	Average
Imorphics	84.36 ± 7.11	87.07 ± 3.36	85.72 ± 5.90
All combined	82.96 ± 8.25	87.70 ± 3.11	85.33 ± 6.68
Top 5 combined	85.38 ± 6.13	87.09 ± 3.22	86.24 ± 5.16
Maximum	87.57 ± 3.37	88.88 ± 1.73	88.23 ± 2.83

in Table 3. Furthermore, statistical analysis on the complete set of case scores was also performed to determine which algorithms are significantly better than other algorithms. As a test repeated measures ANOVA was used in combination with Bonferroni correction at a significance level of 0.05. The results indicated that the top 2 algorithms by Imorphics and ScrAutoProstate are significantly better then every algorithm outside of the top 3. This also holds for both combination strategies. However, none of the algorithms or combinations strategies performed significantly better than the second observer. Finally, the robustness of the algorithms against multi-center data was also tested using ANOVA, but the center did not have a significant impact on the overall algorithm score ($p = 0.118$). The average scores and standard deviations for the algorithms on a per-center basis are presented in Tables 4 and 5.

6. Discussion

6.1. Challenge setup and participation

The images used in the challenge are a good representation of what would be encountered in a clinical setting, with large differences in acquisition protocol, prostate appearance and size. Additionally, the images originated from different centers and scanner manufacturers. The training and test sets were also large enough to draw statistical conclusions on algorithm performance.

The reference standard was constructed by 3 different observers, who each segmented a part of the data. These segmentations were subsequently inspected by the experienced observer for correctness and consistency. Obtaining additional observers for each case would be preferable, however recruiting multiple observers to spend time contouring 100 prostate MR cases is extremely challenging.

The metrics that were used result in a good separation between algorithms and the conversion into per case scores keeps these differences intact. Other metrics were also considered, for example the Jaccard index, sensitivity/specificity and regular Hausdorff distance. Jaccard index is a volume-based metric with similar characteristics as the Dice coefficient, however, in prostate segmentation literature, the Dice coefficient is more often used. To allow better comparison to existing and future literature we chose the Dice coefficient. Sensitivity and specificity are generally not useful in prostate segmentation because specificity will not be very discriminative: the prostate is always a relative small part of the total image volume. Finally, the modified 95% Hausdorff distance was used because the regular Hausdorff distance can be harsh and sensitive to noise: a single pixel can determine overall image segmentation outcome.

One issue with basing case scores on observer reference standards is that very high scores end up in the realm of inter-observer variability. A score higher than 85 is probably still indicative of improved performance, as the second observer segmentations are less accurate than the reference standard, but it is difficult to say whether a score of e.g. 94 is indeed better or just different and equally accurate than a score of 92. However, in general, the algo-

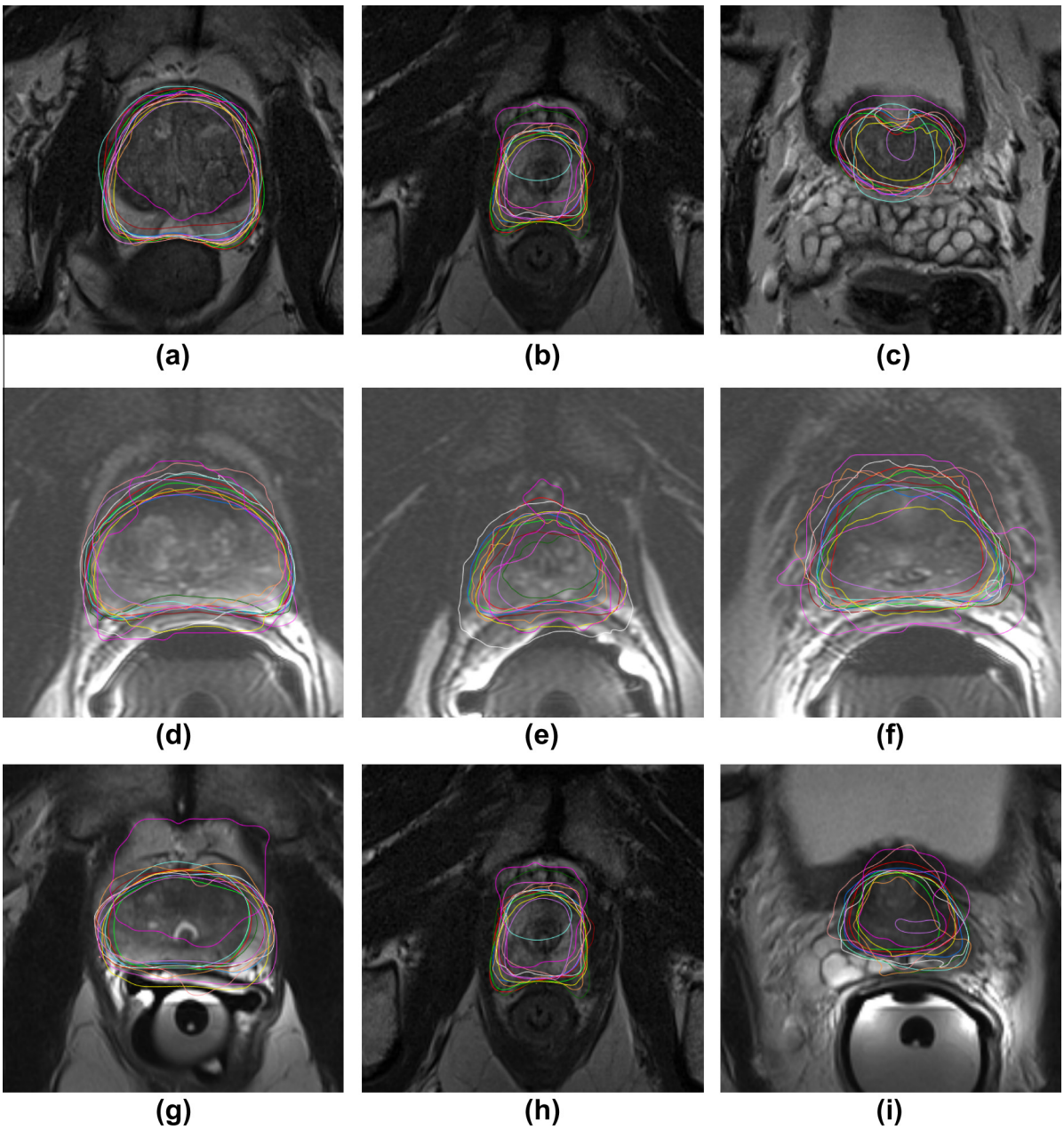


Fig. 5. Qualitative segmentation results of case 3 (a, b, c), case 10 (d, e, f) and case 25 (g, h, i) at the center (a, d, g), apex (b, e, h) and base (c, f, i) of the prostate. Case 3 had the best, case 10 reasonable and case 25 the worst algorithm scores on average. The different colors indicate the results for the different teams. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

rithms in this challenge do not obtain these scores on average, so this is not an issue. Visual inspection of the segmentation results also confirms this, the largest segmentation errors made by the algorithms would not be made by an experienced observer.

An alternative scoring approach that is not sensitive to inter-observer variability is to rank algorithms based on their average rank for each of the sub-scores over all algorithms (e.g. if an algorithm has the highest average Dice of all algorithms, it will have rank 1 for Dice. If the same algorithm has rank 5 for average boundary distance over all algorithms, his average total rank would be 3). This approach has its own disadvantage, e.g. high ranks do not mean good segmentations and algorithm ranking is not only based on the performance of the algorithm itself, but also on the results of other algorithms, i.e. a new algorithm which does very poorly on all metrics except one might influence the ranking of all other algorithms by changing their average rank.

Participation in the initial phase of the challenge was similar to what we have seen in other segmentation challenges, for example (Heimann et al., 2009) and the knee cartilage segmentation challenge (SKI10, http://www.ski10.org). The literature on prostate segmentation is well represented by the competing algorithms, which include active shape models, atlas-based methods, pattern recognition algorithms and variants.

We specifically chose to allow only single submissions per algorithm instead of allowing each group to submit results with different parameter settings, to make sure there would be 'no training on the test set'.

6.2. Challenge results

All algorithms submitted to the challenge produced reasonable to excellent results on average (online and live challenge combined

scores ranging from 68.97 to 85.72). One point to note is that although some algorithms may have received the same average score, the variability can differ substantially, as shown in Tables 6, 7 and 2. For example, the algorithm presented by Robarts (Yuan et al., 2012) scored 77.32 and 80.08 in the online and live challenge respectively, but has a very low variability: 5.51 score standard deviation overall. This is much lower than the algorithms that had similar scores, for Example 7.86 for CBA (Malmberg et al., 2012) and 9.37 for Utwente (Maan and van der Heijden, 2012). Depending on the purpose for which an algorithm is used in the clinic, this can be a very important aspect. As such, it might be good to incorporate performance variability directly in algorithm ranking in future challenges.

It is worth noting that the top 2 algorithms by Imorphics (Vincent et al., 2012) and ScrAutoProstate (Birkbeck et al., 2012) were completely automatic and even outperformed the completely interactive method presented by CBA. Whereas the algorithm by Imorphics performed best overall, the algorithm by ScrAutoProstate should be noted for its exceptionally fast segmentation speed (2.3 s, Table 8), the fastest of all algorithms. Further details about interaction, implementation details and computation time can be found in Table 8. Algorithm computation times varied, with the active shape model based approaches often having computation times in the order of minutes, whereas the atlas based approaches required substantially more time or computing power (e.g. clusters, GPU). It is important to note that some algorithms were implemented in high-level programming languages like Matlab, whereas some where implemented in low-level languages like C++, computation time is thus not only dependent on algorithm efficiency but also on the development platform.

Inspecting the illustrative results in Fig. 5 one can see that algorithms can differ quite substantially per case. In this figure we present the best, worst and a reasonable case with respect to average algorithm performance. Case 25 was especially tricky as it had a large area of fat around the prostate, especially near the base which appears very similar to prostate peripheral zone. Most algorithms oversegmented the prostatic fat, and as the prostate was relatively small, this results in large volumetric errors. However, if one inspects case 25 carefully, it is possible to make the distinction between fat and prostate, especially if you go through the different slices. It is thus no surprise that the interactive segmentation technique of CBA performed the best. Further inspection of the results shows that in the cases with low average algorithm performance the interactive method is usually the best algorithm (e.g. Fig. 3: cases 4, 16 and 21 of the online challenge). This indicates that these cases cause problems for automated methods.

In this challenge we explicitly included segmentation results at the base and the apex of the prostate into the algorithm scoring because these areas are usually the most difficult to segment. This can also be observed in the results, especially Tables 6 and 7. Every algorithm performed worse on the apex and base if we look at the metric values (especially the Dice coefficient and the relative volume difference) themselves; however, as these areas are also the most difficult for the human observer, the scores for apex and base tend to be higher than the overall score. Interesting to note is that the top 2 algorithms outperform the second observer at almost every metric for both apex and base, whereas the overall score is lower than the second observer. For the live challenge the Imorphics algorithm even outperforms the second observer in the overall score. This indicates that for this part of the prostate automatic algorithms might improve over human observers.

Interestingly, similar to the SLIVER07-challenge, active shape based algorithms seemed to give the best results (places 1, 2, 4 and 5), although two of these systems are semi-automatic. Looking at the results in more detail, we can see that the atlas based

Table 4
Average scores and standard deviations per team over the different centers for the online challenge.

	RUNMC	BIDMC	UCL	HK
Imorphics	82.55 ± 8.72	89.05 ± 2.29	84.78 ± 7.52	81.44 ± 7.41
ScrAutoProstate	85.76 ± 3.56	86.26 ± 3.73	83.12 ± 4.95	79.47 ± 8.46
CBA	76.05 ± 7.71	80.82 ± 6.37	83.16 ± 6.17	82.06 ± 4.94
Robarts	77.38 ± 4.73	76.34 ± 5.13	77.57 ± 3.55	77.88 ± 3.77
Utwente	72.52 ± 10.27	78.85 ± 8.11	76.50 ± 13.02	73.16 ± 11.46
Grislies	81.10 ± 9.69	86.10 ± 6.35	77.99 ± 14.82	66.54 ± 10.99
ICProstateSeg	72.70 ± 10.58	82.12 ± 4.71	77.37 ± 7.49	72.40 ± 11.92
DIAG	66.60 ± 13.25	77.48 ± 5.09	81.45 ± 6.76	67.51 ± 20.15
SBIA	81.02 ± 8.77	77.04 ± 10.41	77.31 ± 7.32	78.15 ± 8.19
Rutgers	63.98 ± 14.82	67.00 ± 11.99	69.98 ± 11.02	62.79 ± 16.46
UBUdG	73.17 ± 2.88	67.52 ± 14.90	74.31 ± 6.39	66.73 ± 8.33
Average	75.69 ± 8.63	78.96 ± 7.19	78.50 ± 8.09	73.47 ± 10.19

Table 5
Average scores and standard deviations per team over the different centers for the live challenge. Note that team UBUdG did not participate in the live challenge and as such is not included here.

	RUNMC	BIDMC	UCL	HK
Imorphics	86.86 ± 3.39	88.54 ± 4.17	86.96 ± 3.22	85.92 ± 3.59
ScrAutoProstate	85.06 ± 2.26	85.44 ± 3.03	86.39 ± 3.67	83.44 ± 5.44
CBA	81.32 ± 8.52	83.11 ± 7.88	77.86 ± 16.87	82.53 ± 4.59
Robarts	81.29 ± 5.27	74.77 ± 11.76	81.96 ± 3.99	82.31 ± 5.34
Utwente	80.42 ± 5.48	79.46 ± 7.51	80.64 ± 10.40	80.50 ± 8.42
Grislies	79.98 ± 6.64	77.91 ± 12.30	72.18 ± 16.30	67.33 ± 7.22
ICProstateSeg	82.75 ± 4.67	86.36 ± 3.18	85.60 ± 1.71	80.25 ± 5.83
RUNMC	61.77 ± 15.90	81.51 ± 6.82	83.16 ± 5.35	81.61 ± 3.80
SBIA	79.03 ± 7.21	13.57 ± 30.33	75.50 ± 10.38	77.41 ± 4.39
Rutgers	72.39 ± 14.09	75.10 ± 8.95	65.45 ± 14.78	74.14 ± 6.24
Average	79.09 ± 7.34	74.58 ± 9.59	79.57 ± 8.67	79.54 ± 5.49

Table 6
Averages and standard deviations for all metrics for all teams in the online challenge. Entries indicated with an asterisk had cases with infinite boundary distance measures removed from the average, which could occur due to empty base or apex segmentation results.

Team name	Average boundary distance					
	Overall	Base	Apex	Score (Overall)	Score (Base)	Score (Apex)
Imorphics	2.10 ± 0.68	2.18 ± 1.14	1.96 ± 0.80	82.66 ± 5.60	85.20 ± 7.75	88.44 ± 4.71
ScrAutoProstate	2.13 ± 0.48	2.23 ± 0.70	2.18 ± 0.68	82.42 ± 3.93	84.87 ± 4.73	87.17 ± 3.98
CBA	2.33 ± 0.59	2.60 ± 1.47	2.44 ± 0.81	80.77 ± 4.88	82.31 ± 9.96	85.62 ± 4.75
Robarts	2.65 ± 0.37	2.92 ± 0.88	3.49 ± 0.95	78.09 ± 3.06	80.14 ± 5.97	79.45 ± 5.58
Utwente	3.03 ± 1.06	3.45 ± 1.96	2.68 ± 0.98	74.96 ± 8.73	76.54 ± 13.34	84.20 ± 5.79
Grislies	2.96 ± 1.55	3.19 ± 2.00	2.46 ± 1.26	75.55 ± 12.80	78.35 ± 13.59	85.50 ± 7.42
ICProstateSeg	2.86 ± 0.82	3.18 ± 1.32	2.89 ± 1.05	76.34 ± 6.78	78.38 ± 9.00	82.99 ± 6.21
DIAG	3.40 ± 1.72	4.23 ± 3.06	2.72 ± 1.75	71.90 ± 14.18	71.29 ± 20.81	84.01 ± 10.33
SBIA	2.85 ± 0.72	2.82 ± 1.02	2.13 ± 0.80	76.47 ± 5.94	80.86 ± 6.93	87.44 ± 4.74
Rutgers	4.06 ± 1.80	4.82 ± 2.64*	3.71 ± 1.26*	66.47 ± 14.87	63.06 ± 23.71	74.68 ± 16.56
UBUdG	4.26 ± 1.58	4.21 ± 1.42	4.53 ± 1.71	64.84 ± 13.09	71.40 ± 9.63	73.33 ± 10.08
All combined	2.06 ± 0.78	2.60 ± 1.53	2.04 ± 0.81	82.96 ± 6.46	82.30 ± 10.36	87.98 ± 4.76
Top 5 combined	1.94 ± 0.48	2.10 ± 0.82	1.77 ± 0.62	84.00 ± 3.95	85.70 ± 5.56	89.57 ± 3.63
Maximum	1.78 ± 0.35	1.82 ± 0.52	1.58 ± 0.35	85.28 ± 2.91	87.66 ± 3.51	90.70 ± 2.06
SecondObserver	1.82 ± 0.36	2.21 ± 0.80	2.55 ± 1.08	85.00 ± 2.93	85.00 ± 5.42	85.00 ± 6.34
	95% Hausdorff distance					
Imorphics	5.94 ± 2.14	5.45 ± 2.58	4.73 ± 1.68	84.20 ± 5.70	86.98 ± 6.15	88.84 ± 3.97
ScrAutoProstate	5.58 ± 1.49	5.60 ± 2.35	4.93 ± 1.38	85.15 ± 3.98	86.63 ± 5.62	88.37 ± 3.25
CBA	6.57 ± 2.11	6.64 ± 4.07	5.75 ± 1.91	82.50 ± 5.61	84.15 ± 9.73	86.43 ± 4.52
Robarts	6.48 ± 1.56	6.83 ± 2.26	7.36 ± 2.11	82.76 ± 4.15	83.70 ± 5.39	82.62 ± 4.98
Utwente	7.32 ± 2.44	7.69 ± 3.75	5.89 ± 1.93	80.52 ± 6.48	81.64 ± 8.94	86.11 ± 4.57
Grislies	7.90 ± 3.83	7.61 ± 4.11	5.82 ± 2.82	78.97 ± 10.19	81.85 ± 9.81	86.26 ± 6.65
ICProstateSeg	7.20 ± 1.96	7.27 ± 2.92	6.51 ± 2.31	80.84 ± 5.21	82.64 ± 6.97	84.62 ± 5.46
DIAG	8.59 ± 4.00	9.00 ± 4.62	5.91 ± 3.68	77.15 ± 10.66	78.52 ± 11.04	86.05 ± 8.69
SBIA	7.73 ± 2.68	6.99 ± 2.25	4.60 ± 1.31	79.43 ± 7.14	83.32 ± 5.37	89.14 ± 3.10
Rutgers	9.25 ± 3.76	9.88 ± 4.04*	7.58 ± 2.35*	75.37 ± 10.00	71.18 ± 21.41	78.82 ± 16.23
Rutgers	9.25 ± 3.76	9.88 ± 4.04*	7.58 ± 2.35*	75.37 ± 10.00	71.18 ± 21.41	78.82 ± 16.23
UBUdG	9.17 ± 3.48	9.06 ± 2.71	9.54 ± 3.52	75.59 ± 9.27	78.38 ± 6.46	77.48 ± 8.30
All combined	5.43 ± 2.18	6.00 ± 3.06	4.97 ± 1.94	85.55 ± 5.81	85.67 ± 7.30	88.26 ± 4.57
Top 5 combined	5.30 ± 1.60	5.37 ± 2.38	4.22 ± 1.25	85.91 ± 4.26	87.19 ± 5.67	90.04 ± 2.94
Maximum	4.63 ± 1.06	4.32 ± 1.28	3.67 ± 0.70	87.67 ± 2.81	89.68 ± 3.05	91.34 ± 1.64
SecondObserver	5.64 ± 1.73	6.28 ± 2.95	6.36 ± 2.40	85.00 ± 4.61	85.00 ± 7.04	85.00 ± 5.66
	Dice coefficient					
Imorphics	0.88 ± 0.04	0.86 ± 0.08	0.85 ± 0.08	81.96 ± 6.62	84.76 ± 8.93	88.57 ± 6.13
ScrAutoProstate	0.87 ± 0.04	0.86 ± 0.04	0.83 ± 0.07	81.14 ± 5.39	85.02 ± 4.58	87.79 ± 5.23
CBA	0.87 ± 0.04	0.84 ± 0.07	0.80 ± 0.11	79.80 ± 5.36	82.87 ± 8.07	85.46 ± 7.98
Robarts	0.84 ± 0.03	0.81 ± 0.05	0.71 ± 0.12	75.32 ± 4.25	79.77 ± 5.82	78.70 ± 8.84
Utwente	0.82 ± 0.07	0.78 ± 0.13	0.78 ± 0.09	72.97 ± 9.77	76.12 ± 13.85	84.10 ± 6.44
Grislies	0.83 ± 0.08	0.81 ± 0.11	0.82 ± 0.10	75.10 ± 12.38	79.17 ± 11.85	86.65 ± 7.09
ICProstateSeg	0.82 ± 0.06	0.76 ± 0.13	0.74 ± 0.13	72.68 ± 9.40	74.12 ± 14.15	80.47 ± 9.41
DIAG	0.80 ± 0.09	0.71 ± 0.22	0.79 ± 0.12	69.62 ± 14.20	68.38 ± 23.42	84.82 ± 8.77
SBIA	0.84 ± 0.06	0.81 ± 0.08	0.84 ± 0.07	75.29 ± 8.27	79.29 ± 9.07	88.11 ± 5.31
Rutgers	0.74 ± 0.10	0.61 ± 0.25	0.66 ± 0.17	61.05 ± 15.36	57.75 ± 25.70	74.93 ± 12.60
UBUdG	0.71 ± 0.11	0.71 ± 0.12	0.63 ± 0.14	56.73 ± 16.09	68.17 ± 12.80	72.53 ± 10.20
All combined	0.88 ± 0.05	0.81 ± 0.13	0.81 ± 0.11	81.29 ± 7.55	78.90 ± 14.20	86.31 ± 8.39
Top 5 combined	0.89 ± 0.03	0.87 ± 0.05	0.87 ± 0.06	83.65 ± 4.82	85.79 ± 5.96	90.32 ± 4.63
Maximum	0.90 ± 0.02	0.89 ± 0.03	0.88 ± 0.03	85.08 ± 3.55	88.20 ± 3.80	91.46 ± 2.50
SecondObserver	0.90 ± 0.03	0.86 ± 0.06	0.80 ± 0.11	85.00 ± 3.82	85.00 ± 6.14	85.00 ± 8.39
	Relative volume difference					
Imorphics	2.92 ± 15.71	1.01 ± 19.56	0.65 ± 30.68	72.53 ± 25.31	84.03 ± 16.94	84.20 ± 16.97
ScrAutoProstate	11.53 ± 14.05	9.65 ± 16.52	14.08 ± 34.25	68.18 ± 27.94	82.67 ± 14.82	82.52 ± 18.44
CBA	12.75 ± 13.99	18.85 ± 24.88	0.41 ± 28.63	63.48 ± 25.38	72.51 ± 24.00	82.04 ± 11.91
Robarts	10.31 ± 17.92	12.69 ± 26.26	−3.27 ± 39.09	61.70 ± 28.63	70.65 ± 18.41	74.96 ± 15.61
Utwente	22.30 ± 27.88	27.52 ± 41.86	15.10 ± 41.30	50.19 ± 32.42	57.94 ± 31.74	77.45 ± 23.46
Grislies	19.81 ± 31.93	23.12 ± 44.71	15.46 ± 43.71	59.25 ± 38.47	64.73 ± 31.20	79.31 ± 23.00
ICProstateSeg	−2.61 ± 24.86	−4.47 ± 35.14	−13.31 ± 43.42	57.96 ± 34.16	66.62 ± 25.50	75.09 ± 20.77
DIAG	4.66 ± 28.30	−9.34 ± 43.13	11.66 ± 54.14	51.04 ± 31.02	60.62 ± 31.86	76.15 ± 24.37
SBIA	16.19 ± 25.35	13.47 ± 30.78	11.26 ± 35.57	51.63 ± 35.95	67.71 ± 23.49	81.33 ± 21.19
Rutgers	−5.83 ± 30.81	−22.11 ± 57.39	−16.68 ± 46.37	52.18 ± 30.04	44.52 ± 31.99	71.58 ± 24.00
UBUdG	−5.16 ± 21.40	−7.33 ± 28.05	−14.55 ± 33.25	59.02 ± 24.71	69.96 ± 16.63	77.87 ± 16.16
All combined	−10.02 ± 14.62	−15.45 ± 25.94	−19.44 ± 22.45	67.17 ± 25.33	73.19 ± 23.89	81.67 ± 13.00
Top 5 combined	7.63 ± 13.45	7.32 ± 18.53	6.37 ± 27.31	73.70 ± 25.02	82.15 ± 15.60	86.50 ± 16.37
Maximum	2.76 ± 3.05	4.50 ± 4.80	4.23 ± 4.21	93.48 ± 7.19	94.61 ± 5.76	96.78 ± 3.21
SecondObserver	−1.87 ± 7.32	−6.17 ± 13.49	−16.24 ± 21.13	85.00 ± 9.23	85.00 ± 9.23	85.00 ± 13.57

systems comparatively have more trouble with cases which are not well represented by the training set, for example case 23, which has a prostate volume of 325 mL, while the average is around 50 mL.

One interactive method was included (team CBA) which on average scored 80.94, which is considerably lower than the second observer. This is mostly caused by over-segmentation at the base of the prostate, often the seminal vesicles were included in the

Table 7
Averages and standard deviations for all metrics for all teams in the live challenge. Entries indicated with an asterisk had cases with infinite boundary distance measures removed from the average, which could occur due to empty segmentation results.

Team name	Average boundary distance					
	Overall	Base	Apex	Score (Overall)	Score (Base)	Score (Apex)
Imorphics	1.95 ± 0.36	2.45 ± 0.65	1.83 ± 0.53	85.53 ± 2.70	87.12 ± 3.41	88.21 ± 3.39
ScrAutoProstate	2.18 ± 0.36	2.34 ± 0.78	2.16 ± 0.70	83.86 ± 2.65	87.73 ± 4.12	86.05 ± 4.50
CBA	2.56 ± 0.96	2.48 ± 1.55	119.28 ± 522.54	81.03 ± 7.10	86.95 ± 8.13	80.06 ± 21.12
Robarts	2.67 ± 0.62	2.66 ± 0.90	3.93 ± 2.42	80.23 ± 4.56	86.01 ± 4.74	74.64 ± 15.57
Utwente	2.87 ± 0.79	3.47 ± 1.33	2.43 ± 0.72	78.79 ± 5.87	81.76 ± 6.96	84.32 ± 4.62
Grislies	4.17 ± 2.35	3.75 ± 2.25	2.82 ± 1.06	69.14 ± 17.43	80.31 ± 11.80	81.81 ± 6.84
ICProstateSeg	2.35 ± 0.99	2.62 ± 1.37	1.95 ± 0.96	82.63 ± 7.35	86.23 ± 7.21	87.46 ± 6.16
DIAG	3.21 ± 1.39	237.53 ± 718.80	2.31 ± 0.71	76.26 ± 10.29	71.09 ± 27.15	85.11 ± 4.57
SBIA	3.13 ± 0.74*	3.13 ± 0.64*	2.89 ± 1.03*	61.49 ± 31.92	66.83 ± 34.41	65.10 ± 33.91
Rutgers	3.84 ± 1.37	3.70 ± 1.12*	4.21 ± 1.83	71.54 ± 10.18	72.52 ± 25.41	72.87 ± 11.80
All combined	1.97 ± 0.34	2.18 ± 0.64	1.82 ± 0.53	85.43 ± 2.51	88.55 ± 3.35	88.28 ± 3.41
Top 5 combined	1.90 ± 0.32	2.15 ± 0.80	1.92 ± 0.64	85.93 ± 2.37	88.70 ± 4.18	87.61 ± 4.14
Maximum	1.87 ± 0.30	1.82 ± 0.45	1.53 ± 0.30	86.17 ± 2.20	90.44 ± 2.36	90.17 ± 1.88
SecondObserver	2.03 ± 0.50	2.86 ± 1.26	2.33 ± 1.35	85.00 ± 3.73	85.00 ± 6.63	85.00 ± 8.69
	95% Hausdorff distance					
Imorphics	5.54 ± 1.74	6.09 ± 1.61	4.58 ± 1.36	86.35 ± 4.28	87.96 ± 3.19	87.03 ± 3.86
ScrAutoProstate	6.04 ± 1.67	5.64 ± 2.17	4.60 ± 1.39	85.11 ± 4.12	88.84 ± 4.29	86.96 ± 3.94
CBA	7.34 ± 3.08	6.29 ± 3.03	122.28 ± 523.16	81.90 ± 7.59	87.55 ± 6.00	80.72 ± 20.05
Robarts	7.15 ± 2.08	6.12 ± 2.14	7.76 ± 3.20	82.38 ± 5.12	87.89 ± 4.22	78.01 ± 9.06
Utwente	6.72 ± 1.42	7.42 ± 2.38	5.68 ± 1.66	83.43 ± 3.51	85.33 ± 4.71	83.91 ± 4.70
Grislies	11.08 ± 5.85	8.68 ± 4.61	6.88 ± 2.21	72.68 ± 14.42	82.83 ± 9.11	80.49 ± 6.27
ICProstateSeg	5.89 ± 2.59	5.64 ± 2.73	4.58 ± 2.35	85.48 ± 6.38	88.83 ± 5.41	87.00 ± 6.67
DIAG	7.95 ± 3.21	242.13 ± 719.85	4.74 ± 1.34	80.40 ± 7.91	75.30 ± 26.69	86.56 ± 3.79
SBIA	7.07 ± 1.64*	7.21 ± 1.96*	5.93 ± 1.69*	66.05 ± 34.07	68.59 ± 35.35	66.54 ± 34.40
Rutgers	8.48 ± 2.53	242.00 ± 719.42	7.82 ± 2.42	79.09 ± 6.23	75.29 ± 26.10	77.82 ± 6.86
All combined	5.67 ± 1.82	5.14 ± 1.40	4.46 ± 1.46	86.01 ± 4.49	89.84 ± 2.78	87.35 ± 4.13
Top 5 combined	5.49 ± 1.54	5.48 ± 2.24	4.56 ± 1.51	86.45 ± 3.80	89.16 ± 4.43	87.07 ± 4.27
Maximum	4.80 ± 1.02	4.20 ± 0.94	3.53 ± 0.76	88.17 ± 2.52	91.69 ± 1.86	90.13 ± 2.10
SecondObserver	6.08 ± 2.23	7.58 ± 3.90	5.29 ± 2.53	85.00 ± 5.50	85.00 ± 7.71	85.00 ± 7.17
	Dice coefficient					
Imorphics	0.89 ± 0.03	0.84 ± 0.06	0.86 ± 0.07	85.51 ± 3.92	86.98 ± 5.21	89.15 ± 5.66
ScrAutoProstate	0.87 ± 0.03	0.85 ± 0.06	0.83 ± 0.10	83.17 ± 3.53	87.35 ± 5.20	86.81 ± 7.47
CBA	0.85 ± 0.08	0.85 ± 0.10	0.77 ± 0.23	79.69 ± 10.77	87.82 ± 8.16	82.13 ± 17.39
Robarts	0.84 ± 0.04	0.84 ± 0.06	0.67 ± 0.22	78.82 ± 5.40	86.62 ± 4.90	74.31 ± 17.27
Utwente	0.83 ± 0.06	0.77 ± 0.10	0.79 ± 0.10	77.46 ± 7.61	81.40 ± 7.81	84.12 ± 7.47
Grislies	0.77 ± 0.12	0.78 ± 0.12	0.79 ± 0.09	70.04 ± 16.09	81.93 ± 9.79	83.82 ± 7.30
ICProstateSeg	0.76 ± 0.26	0.72 ± 0.26	0.74 ± 0.26	71.70 ± 25.03	77.24 ± 21.30	80.26 ± 20.36
DIAG	0.80 ± 0.07	0.63 ± 0.30	0.82 ± 0.07	73.81 ± 9.43	69.73 ± 24.09	86.18 ± 5.71
SBIA	0.65 ± 0.34	0.64 ± 0.34	0.63 ± 0.33	60.41 ± 31.93	70.99 ± 27.41	71.78 ± 25.83
Rutgers	0.75 ± 0.10	0.68 ± 0.25	0.62 ± 0.22	67.41 ± 13.75	73.93 ± 20.13	70.85 ± 17.08
All combined	0.89 ± 0.03	0.87 ± 0.05	0.86 ± 0.08	86.10 ± 3.30	89.01 ± 4.10	88.93 ± 5.88
Top 5 combined	0.89 ± 0.02	0.87 ± 0.06	0.85 ± 0.09	86.12 ± 2.90	89.03 ± 4.94	88.21 ± 6.58
Maximum	0.90 ± 0.02	0.89 ± 0.03	0.89 ± 0.03	86.51 ± 2.47	90.97 ± 2.82	91.90 ± 1.97
SecondObserver	0.89 ± 0.03	0.82 ± 0.10	0.81 ± 0.15	85.00 ± 4.18	85.00 ± 8.32	85.00 ± 11.56
	Relative volume difference					
Imorphics	−1.50 ± 9.15	−8.31 ± 18.08	−1.03 ± 23.97	86.31 ± 13.01	87.15 ± 7.70	87.55 ± 10.37
ScrAutoProstate	10.05 ± 11.56	7.77 ± 22.01	9.59 ± 30.51	73.96 ± 17.56	86.55 ± 11.38	84.60 ± 15.29
CBA	12.26 ± 17.73	24.75 ± 41.69	−7.05 ± 39.63	63.49 ± 24.70	81.63 ± 24.91	81.50 ± 20.08
Robarts	−1.72 ± 17.47	5.30 ± 25.52	−29.19 ± 37.14	71.84 ± 21.87	86.46 ± 14.29	73.77 ± 18.61
Utwente	12.62 ± 22.25	20.75 ± 37.43	0.66 ± 28.70	62.15 ± 30.81	75.02 ± 20.62	85.40 ± 12.77
Grislies	43.13 ± 65.32	36.41 ± 58.73	7.23 ± 38.19	37.72 ± 40.30	72.42 ± 29.35	79.01 ± 15.76
ICProstateSeg	−8.49 ± 34.17	−14.15 ± 34.88	−14.88 ± 36.55	69.10 ± 29.32	81.82 ± 22.02	80.77 ± 18.68
DIAG	−12.34 ± 18.38	−38.10 ± 32.87	1.61 ± 28.65	64.59 ± 25.81	70.54 ± 24.65	84.60 ± 11.74
SBIA	6.55 ± 59.45	2.66 ± 57.32	12.12 ± 68.31	30.65 ± 34.34	64.84 ± 24.89	62.50 ± 28.06
Rutgers	−14.59 ± 26.52	−24.79 ± 31.88	−24.37 ± 47.01	50.76 ± 27.17	76.87 ± 20.18	72.31 ± 22.95
All combined	2.69 ± 9.75	−0.16 ± 13.09	−2.25 ± 24.49	83.77 ± 12.66	91.64 ± 5.14	87.48 ± 10.93
Top 5 combined	4.69 ± 9.95	6.89 ± 20.16	−3.07 ± 26.74	82.19 ± 13.65	88.57 ± 11.35	86.10 ± 11.74
Maximum	1.80 ± 1.43	3.65 ± 3.24	3.58 ± 3.99	96.28 ± 2.94	97.21 ± 2.47	97.52 ± 2.74
SecondObserver	−5.72 ± 7.44	−17.49 ± 18.12	−17.97 ± 22.90	85.00 ± 12.07	85.00 ± 11.93	85.00 ± 13.03

prostate segmentation. Thus this algorithm is very dependent on the operator; in principle the algorithm should be able to get close to expert performance given an expert reader.

There were several semi-automatic algorithms (teams Robarts, UTwente and UBUdG) which needed manual interaction to initialize the algorithms. The interaction types and the influence this interaction has on segmentation accuracy will differ between the algorithms. Although none of the teams have explicitly tested the robustness to different initializations, some general comments can be made. For the Robarts algorithm a number of points on the prostate boundary have to be set (8 to 10) to initialize a shape and the initial foreground and background distributions. As such, the algorithm is robust to misplacing single points. For the Utwente algorithm, the prostate center has to be indicated to

Table 8
Details on computation time, interaction and computer systems used for the different algorithms. If algorithms where multi-threaded (MT) or used the GPU this is also indicated.

Team name	Avg. time	System	MT	GPU	Availability	Remarks
Imorphics	8 min	2.83 GHz 4-cores	No	No	Commercially available (http://www.imorphics.com/).	
ScrAutoProstate	2.3 s	2.7GGz 12-cores	Yes	Not available	No	
CBA	4 min	2.7 GHz 2-cores	No	No	Binaries available at: http://www.cb.uu.se/filip/SmartPaint/	Fully interactive painting
Robarts	45 s	3.2 GHz 1-core, 512 CUDA-cores	No	Yes	Available at http://www.mathworks.com/matlabcentral/fileexchange/34126-fast-continuous-max-flow-algorithm-to-2d3d-image-segmentation	User indicates 8 to 10 points on prostate surface
Utwente	4 min	2.94 GHz 4-cores	Yes	No	Not available	User indicates prostate center
Grislies	7 min	2.5 GHz 4-cores	No	No	Not available	
ICProstateSeg	30 min	3.2 GHz 4-cores, 96 CUDA-cores	No	Yes	Not available	
DIAG	22 min	2.27 GHz 8-cores	No	No	Registration algorithm available on http://elastix.isi.uu.nl/	Runs algorithm on a cluster of 50 cores, average time without cluster 7 min per atlas
SBIA	40 min	2.9 GHz, 2 cores	No	No	Registration algorithm available on http://www.rad.upenn.edu/sbia/software/dramms/	Runs algorithm on a cluster of 140 cores, average time without cluster 25 min per atlas
Rutgers	3 min	2.67 GHz, 8-cores	Yes	No	Not available	
UBUdG	100 s	3.2 GHz 4-cores	No	No	Not available	User selects first and last prostate slice

initialize the active appearance and shape models. Big deviations in point selection can cause problems for active appearance and shape models, however in general they are pretty robust against small deviations (Cootes et al., 2001). For the UBUdG method, the user has to select the first and last slice of the prostate. As such, the algorithm will be unable to segment the prostate if it extends beyond those slices, which is an issue if users cannot correctly identify the start and end slice of the prostate.

Another aspect which plays a role in this challenge was the robustness of the algorithms to multi-center data. The image differences between the centers were actually quite large, especially between the endorectal coil and nonendorectal coil cases, as can be seen in Fig. 1. Differences include coil artifacts near the peripheral zone, coil profiles, image intensities, slice thickness and resolution. However, if we look at for example Tables 4, 5, 7 and 8 and Fig. 3, it can be seen that all submitted algorithms are at least reasonably robust against these differences. We could not find any significant differences in the performance of the algorithms relative to the different centers using ANOVA ($p = 0.118$).

We also investigated whether segmentation performance could be improved by making several algorithm combinations. First, a majority voting on the segmentation results of all algorithms and the top 5 best performing was calculated. Second, to get a reference for the best possible combination we took the best performing score per case. The summary results of these combinations can be found in Table 3. Taking the best results per case results in a substantially better average score than the best performing algorithms. This might be an indication that certain cases might be better suited to some algorithms, and as such, that algorithm selection should be performed on a case-by-case basis. The combinations of algorithms using majority voting also shows that given the correct combination, algorithm results can be improved (84.36 to 85.38 for the online challenge and 87.07 to 87.70 for the live challenge). Although the increase in score is small, it is accompanied by a reduction of the standard deviation (for the top 5 combination strategy, Table 3), as the improvements especially occur in poor performing cases. These scores and the reduction in standard deviation thus show that combining algorithms might result in more robust segmentation. These scores also show that there still is room for improvement for the individual algorithms. How to combine and which algorithms to combine is a nontrivial problem and warrants further investigation.

Finally, to assess the statistical significance of differences in algorithm performance we used repeated measures ANOVA with Bonferroni correction. The methods by Imorphics and ScrAutoProstate perform significantly better than all the algorithms outside of the top 3 ($p < 0.05$).

7. Future work and concluding remarks

Although in general the segmentation algorithms, especially the top 2, gave good segmentation results, some challenges still remain. As we could see in case 25 (Fig. 5), algorithms sometimes struggle with the interface between the prostate and surrounding tissue. This is not only true for peri-prostatic fat, but also for the interface between the prostate and the rectum, the bladder and the seminal vesicles. Part of these challenges could be addressed by increasing through-plane resolution, but integration of these structures into the segmentation algorithms might also improve performance. Examples included coupled active appearance models (Cootes et al., 2000) or hierarchical segmentation strategies (Wolz et al., 2012). Furthermore, the enormous volume differences that can occur in the prostate can also be problematic: case 23 had a volume which was approximately 6 times as large as the average. Automatically selecting appropriate atlas sets or appearance models based on an initial segmentation could be a solution. In the

difficult cases the interactive segmentation method of team CBA was often the best. This shows that automated performance could still be improved.

Future work on prostate segmentation might also focus on the segmentation of related prostatic structures or substructures. Examples are segmentation of the prostatic zones (transition, central and peripheral), the neurovascular bundles or the seminal vesicles.

Solving these remaining issues might lead to algorithms which, for any case, can replace the tedious task of manually outlining by humans without any intervention. Until we are at that level, the challenge itself will remain online for new submissions and can thus be used as a reference for algorithm performance on multi-center data. As such it could lead to more transparency in medical image analysis.

Acknowledgments

This research was funded by Grant KUN2007-3971 from the Dutch Cancer Society and by the National Cancer Institute of the National Institutes of Health under Award Nos. R01CA136535-01, R01CA140772-01, and R21CA167811-01; the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award No. R43EB015199-01; the National Science Foundation under Award No. IIP-1248316; the QED award from the University City Science Center and Rutgers University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Barentsz, J.O., Richenberg, J., Clements, R., Choyke, P., Verma, S., Villeirs, G., Rouviere, O., Logager, V., Fütterer, J.J., European Society of Urogenital Radiology, 2012. ESUR prostate MR guidelines 2012. Eur. Radiol. 22, 746–757.

Bezdek, J.C., Hall, L.O., Clarke, L.P., 1993. Review of MR image segmentation techniques using pattern recognition. Med. Phys. 20, 1033–1048.

Birkbeck, N., Zhang, J., Requardt, M., Kiefer, B., Gall, P., Kevin Zhou, S., 2012. Region-specific hierarchical segmentation of MR prostate using discriminative learning. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Chandra, S.S., Dowling, J.A., Shen, K.K., Raniga, P., Pluim, J.P.W., Greer, P.B., Salvado, O., Fripp, J., 2012. Patient specific prostate segmentation in 3-D magnetic resonance images. IEEE Trans. Med. Imaging 31, 1955–1964.

Clarke, L.P., Velthuizen, R.P., Phuphanich, S., Schellenberg, J.D., Arrington, J.A., Silbiger, M., 1993. MRI: stability of three supervised segmentation techniques. Magn. Reson. Imaging 11, 95–106.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23, 681–685.

Cootes, T.F., Twining, C.J., Petrovic, V., Schestowitz, R., Taylor, C.J., 2005. Groupwise construction of appearance models using piece-wise affine deformations. In: Proceedings of 16th British Machine Vision Conference, pp. 879–888.

Cootes, T.F., Walker, K., Taylor, C.J., 2000. View-based active appearance models. In: Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition. IEEE Comput. Soc., pp. 227–232.

Costa, M.J., Delingette, H., Novellas, S., Ayache, N., 2007. Automatic segmentation of bladder and prostate using coupled 3D deformable models. Med. Image Comput. Comput. Assist. Interv. 10, 252–260.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54, 940–954.

Dickinson, L., Ahmed, H.U., Allen, C., Barentsz, J.O., Carey, B., Fütterer, J.J., Heijmink, S.W., Hoskin, P.J., Kirkham, A., Padhani, A.R., Persad, R., Puech, P., Punwani, S., Sohaib, A.S., Tombal, B., Villers, A., van der Meulen, J., Emberton, M., 2011. Magnetic resonance imaging for the detection, localisation, and characterisation of prostate cancer: recommendations from a european consensus meeting. Eur. Urol. 59, 477–494.

Fütterer, J.J., Barentsz, J., 2009. 3T MRI of prostate cancer. Appl. Radiol. 38, 25–37.

Gao, Q., Rueckert, D., Edwards, P., 2012. An automatic multi-atlas based prostate segmentation using local appearance-specific atlases and patch-based voxel weighting. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Gao, Y., Liao, S., Shen, D., 2012b. Prostate segmentation by sparse representation based classification. Med. Phys. 39, 6372–6387.

Ghose, S., Mitra, J., Oliver, A., Martí, R., Lladó, X., Freixenet, J., Vilanova, J.C., Sidibé, D., Meriaudeau, F., 2012. A random forest based classification approach to prostate segmentation in MRI. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Hambrock, T., Hoeks, C., Hulsbergen-van de Kaa, C., Scheenen, T., Fütterer, J., Bouwense, S., van Oort, I., Schröder, F., Huisman, H., Barentsz, J., 2012. Prospective assessment of prostate cancer aggressiveness using 3-T diffusion-weighted magnetic resonance imaging-guided biopsies versus a systematic 10-core transrectal ultrasound prostate biopsy cohort. Eur. Urol. 61, 177–184.

Heimann, T., van Ginneken, B., Styner, M., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., Bello, F., Binnig, G., Bischof, H., Bornik, A., Cashman, P., Chi, Y., Cordova, A., Dawant, B., Fidrich, M., Furst, J., Furukawa, D., Grenacher, L., Hornegger, J., Kainmuller, D., Kitney, R., Kobatake, H., Lamecker, H., Lange, T., Lee, J., Lennon, B., Li, R., Li, S., Meinzer, H.P., Nemeth, G., Raicu, D., Rau, A.M., van Rikxoort, E., Rousson, M., Rusko, L., Saddi, K., Schmidt, G., Seghers, D., Shimizu, A., Slagmolen, P., Sorantin, E., Soza, G., Susomboon, R., Waite, J., Wimmer, A., Wolf, I., 2009. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans. Med. Imaging 28, 1251–1265.

Hoeks, C.M.A., Hambrock, T., Yakar, D., Hulsbergen-van de Kaa, C.A., Feuth, T., Witjes, J.A., Fütterer, J.J., Barentsz, J.O., 2013. Transition zone prostate cancer: detection and localization with 3-t multiparametric MR imaging. Radiology 266, 207–217.

Hu, Y., Ahmed, H.U., Taylor, Z., Allen, C., Emberton, M., Hawkes, D., Barratt, D., 2012. MR to ultrasound registration for image-guided prostate interventions. Med. Image Anal. 16, 687–703.

Kirschner, M., Jung, F., Wesarg, S., 2012. Automatic prostate segmentation in MR images with a probabilistic active shape model. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Kitajima, K., Kaji, Y., Fukabori, Y., Yoshida, K., Suganuma, N., Sugimura, K., 2010. Prostate cancer detection with 3 T MRI: comparison of diffusion-weighted imaging and dynamic contrast-enhanced MRI in combination with T2-weighted imaging. J. Magn. Reson. Imaging 31, 625–631.

Klein, S., van der Heide, U.A., Lips, I.M., van Vulpen, M., Staring, M., Pluim, J.P.W., 2008. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. Med. Phys. 35, 1407–1417.

Kroon, D., Kowalski, P., Tekieli, W., Reeuwijk, E., Saris, D., Slump, C.H., 2012. MRI based knee cartilage assessment. In: Medical Imaging, pp. 83151V-1–10.

Langerak, T.R., van der Heide, U.A., Kotte, A.N.T.J., Viergever, M.A., van Vulpen, M., Pluim, J.P.W., 2010. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans. Med. Imaging 29, 2000–2008.

Leemput, K.V., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imaging 18, 885–896.

Li, C., Xu, C., Anderson, A.W., Gore, J.C., 2009. MRI tissue classification and bias field estimation based on coherent local intensity clustering: a unified energy minimization framework. In: Inf. Process Med. Imaging, pp. 288–299.

Li, X., Huang, W., Rooney, W.D., 2012. Signal-to-noise ratio, contrast-to-noise ratio and pharmacokinetic modeling considerations in dynamic contrast-enhanced magnetic resonance imaging. Magn. Reson. Imaging 30, 1313–1322.

Litjens, G.J.S., Karssemeijer, N., Huisman, H.J., 2012. A multi-atlas approach for prostate segmentation in MR images. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Lorensen, W.E., Cline, H.E., 1987. Marching cubes: a high resolution 3D surface construction algorithm. In: Computer Graphics: SIGGRAPH '87 Conference Proceedings, pp. 163–169.

Maan, B., van der Heijden, F., 2012. Prostate MR image segmentation using 3D active appearance models. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Makni, N., Puech, P., Lopes, R., Dewalle, A.S., Colot, O., Betrouni, N., 2009. Combining a deformable model and a probabilistic framework for an automatic 3d segmentation of prostate on MRI. Int. J. Comput. Assist. Radiol. Surg. 4, 181–188.

Malmberg, F., Strand, R., Kullberg, J., Nordenskjöld, R., Bengtsson, E., 2012. Smart paint a new interactive segmentation method applied to MR prostate segmentation. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Murphy, K., van Ginneken, B., Reinhardt, J.M., Kabus, S., Ding, K., Deng, X., Cao, K., Du, K., Christensen, G.E., Garcia, V., Vercauteren, T., Ayache, N., Commowick, O., Malandain, G., Glocker, B., Paragios, N., Navab, N., Gorbunova, V., Sporring, J., de Bruijne, M., Han, X., Heinrich, M.P., Schnabel, J.A., Jenkinson, M., Lorenz, C., Modat, M., McClelland, J.R., Ourselin, S., Muenzing, S.E.A., Viergever, M.A., Nigris, D.D., Collins, D.L., Arbel, T., Peroni, M., Li, R., Sharp, G.C., Schmidt-Richberg, A., Ehrhardt, J., Werner, R., Smeets, D., Loeckx, D., Song, G., Tustison, N., Avants, B., Gee, J.C., Staring, M., Klein, S., Stoel, B.C., Urschler, M., Werlberger, M., Vandemeulebroucke, J., Rit, S., Sarrut, D., Pluim, J.P.W., 2011. Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge. IEEE Trans. Med. Imaging 31, 1901–1920.

Niemeijer, M., van Ginneken, B., Cree, M.J., Mizutani, A., Quellec, G., Sánchez, C.I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., Wu, X., Cazuguel, G., You, J., Mayo, A., Li, Q., Hatanaka, Y., Cochener, B., Roux, C., Karray, F., Garcia, M., Fujita, H., Abràmoff, M.D., 2010. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans. Med. Imaging 29, 185–195.

Nyúl, L.G., Udupa, J.K., 1999. On standardizing the MR image intensity scale. Magn. Reson. Med. 42, 1072–1081.

Nyúl, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging 19, 143–150.

Ou, Y., Doshi, J., Erus, G., Davatzikos, C., 2012. Multi-Atlas segmentation of the prostate: a zooming process with robust registration and Atlas selection. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Ou, Y., Sotiras, A., Paragios, N., Davatzikos, C., 2010. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. Med. Image Anal. 15, 622–639.

Pasquier, D., Lacornerie, T., Vermandel, M., Rousseau, J., Lartigau, E., Betrouni, N., 2007. Automatic segmentation of pelvic structures from magnetic resonance images for prostate cancer radiotherapy. Int. J. Radiat. Oncol. Biol. Phys. 68, 592–600.

Pérez, P., Gangnet, M., Blake, A., 2003. Poisson image editing. ACM Trans. Graph 22, 313–318.

Schaap, M., Metz, C.T., van Walsum, T., van der Giessen, A.G., Weustink, A.C., Mollet, N.R., Bauer, C., Bogunović, H., Castro, C., Deng, X., Dikici, E., O'Donnell, T., Frenay, M., Friman, O., Hernández-Hoyos, M., Kitslaar, P.H., Krissian, K., Kühnel, C., Luengo-Oroz, M.A., Orkisz, M., Smedby, O., Styner, M., Szymczak, A., Tek, H., Wang, C., Warfield, S.K., Zambal, S., Zhang, Y., Krestin, G.P., Niessen, W.J., 2009. Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. Med. Image Anal. 13, 701–714.

Shattuck, D.W., Prasad, G., Mirza, M., Narr, K.L., Toga, A.W., 2009. Online resource for validation of brain segmentation methods. Neuroimage 45, 431–439.

Sung, K., Daniel, B.L., Hargreaves, B.A., 2013. Transmit B1+ field inhomogeneity and T(1) estimation errors in breast DCE-MRI at 3 tesla. J. Magn. Reson. Imaging.

Tanimoto, A., Nakashima, J., Kohno, H., Shinmoto, H., Kuribayashi, S., 2007. Prostate cancer screening: the clinical value of diffusion-weighted imaging and dynamic MR imaging in combination with T2-weighted imaging. J. Magn. Reson. Imaging 25, 146–152.

Tiwari, P., Kurhanewicz, J., Madabhushi, A., 2013. Multi-kernel graph embedding for detection, gleason grading of prostate cancer via MRI/mrs. Med. Image Anal. 17, 219–235.

Toth, R., Bloch, B.N., Genega, E.M., Rofsky, N.M., Lenkinski, R.E., Rosen, M.A., Kalyanpur, A., Pungavkar, S., Madabhushi, A., 2011a. Accurate prostate volume estimation using multifeature active shape models on T2-weighted MRI. Acad. Radiol. 18, 745–754.

Toth, R., Madabhushi, A., 2012. Deformable landmark-free active appearance models: application to segmentation of multi-institutional prostate MRI data. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Toth, R., Madabhushi, A., 2012b. Multifeature landmark-free active appearance models: application to prostate MRI segmentation. IEEE Trans. Med. Imaging 31, 1638–1650.

Toth, R., Tiwari, P., Rosen, M., Reed, G., Kurhanewicz, J., Kalyanpur, A., Pungavkar, S., Madabhushi, A., 2011b. A magnetic resonance spectroscopy driven initialization scheme for active shape model based prostate segmentation. Med. Image Anal. 15, 214–225.

Villeirs, G.M., Meerleer, G.O.D., Visschere, P.J.D., Fonteyne, V.H., Verbaeys, A.C., Oosterlinck, W., 2011. Combined magnetic resonance imaging and spectroscopy in the assessment of high grade prostate carcinoma in patients with elevated PSA: a single-institution experience of 356 patients. Eur. J. Radiol. 77, 340–345.

Vincent, G., Guillard, G., Bowes, M., 2012. Fully automatic segmentation of the prostate using active appearance models. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Viola, P., Jones, M., 2001. Robust real-time object detection. Int. J. Comput. Vis..

Vos, P.C., Barentsz, J.O., Karssemeijer, N., Huisman, H.J., 2012. Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis. Phys. Med. Biol. 57, 1527–1542.

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

Wolz, R., Chu, C., Misawa, K., Mori, K., Rueckert, D., 2012. Multi-organ abdominal CT segmentation using hierarchically weighted subject-specific atlases. Med. Image Comput. Comput. Assist. Interv. 15, 10–17.

Yuan, J., Qiu, W., Ukwatta, E., Rajchl, M., Sun, Y., Fenster, A., 2012. An efficient convex optimization approach to 3D prostate MRI segmentation with generic star shape prior. MICCAI Grand Challenge: Prostate MR Image Segmentation 2012.

Zheng, B., Tan, J., Ganott, M.A., Chough, D.M., Gur, D., 2009. Matching breast masses depicted on different views: a comparison of three methods. Acad. Radiol. 16, 1338–1347.