Automatic Segmentation of Abdominal MRI Using Selective Sampling and Random Walker

Janine Thoma[1], Firat Ozdemir[2], and Orcun Goksel[2]

[1] Computer Vision Lab, ETH Zurich, Switzerland
[2] Computer-assisted Applications in Medicine, ETH Zurich, Switzerland

Abstract. MRI segmentation is a challenging task due to low anatomical contrast and large inter-patient variation. We propose a feature-driven automatic segmentation framework, combining voxel-wise classification with a Random-Walker (RW) based spatial regularization. Typically, such steps are treated independently, i.e. classification outcome is maximized without taking into account the regularization to follow. Herein we present a method for selective sampling of training patches, in view of the posterior spatial regularization. This aims to concentrate training samples near desired anatomical boundaries, around which the gain from a subsequent RW regularization will potentially be minimal. This trades off a lower classification accuracy for a higher joint segmentation performance. We compare our proposed sampling strategy to conventional uniform sampling on 20 full-body MR T1 scans from the VISCERAL dataset, both with RW and Markov Random Fields regularizations, showing Dice improvements of up to 12× with the proposed approach.

1 Introduction

Segmentation of abdominal organs in medical images plays an important role in therapy planning and diagnosis. In the clinics, manual segmentation is currently still a common practice; leading to long processing times, subjectivity in the resulting segmentations, and high time and cost expenditure of trained physicians. Therefore, there is a need and significant interest for reliable automatic segmentation methods. In recent years a number of multi-organ segmentation methods have achieved promising results for both contrast-enhanced and non-contrast-enhanced computed tomography (CT), e.g. [1–7]. The majority of these methods use a form of multi-atlas segmentation [1, 2, 8, 4, 6]. Alternatively, in [5] multi-boost learning and statistical shape models are used, and in [7] active appearance models, live-wires and graph-cuts are used.

Unlike CT, magnetic resonance imaging (MRI) does not use ionizing radiation for data acquisition, making it an attractive alternative where applicable. Its segmentation, however, is a challenging task due to high variability in anatomy appearance, low contrast across structures, and large inter-patient and inter-scan variation. Thus, state-of-the-art in abdominal MRI segmentation is also

often significantly inferior to that of CT segmentation. In [9], an MRI multi-organ segmentation method is presented by combining kernel graph-cuts with shape-priors. The multi-atlas segmentation methods in [3] and [4] are invariant to image modality and thus have been evaluated also on MR images. Nevertheless, these results for MRI are not satisfactory for most clinical applications.

Typically, methods developed for MRI segmentation are optimized for one specific abdominal organ, structure, or condition. One example of specialized segmentation is [10], which proposes a method for renal compartment segmentation. Another example is the unsupervised myocardial segmentation presented in [11]. Some of these methods exploit the characteristics of a specific organ appearance or MR acquisition type such as DCE-MRI in [10] or CINE/CP-BOLD in [11]. Nevertheless, there are not many techniques that apply to the problem of multi-organ abdominal MRI segmentation with high accuracy.

In this paper, we propose an automatic multi-organ segmentation method for abdominal organs, which we test on unenhanced full-body T1 MRI from the VISCERAL Anatomy organ segmentation dataset [12]. Our method consists of random forest (RF) [13] based voxel classification followed by random walker (RW) based spatial regularization. The main contribution of this paper lies in the holistic analysis of a selective sampling strategy for classifier training, considering the subsequent spatial regularization.

2 Methods

2.1 Features

For training and applying the classification, we extract image features that encode absolute normalized positions, statistical properties, anisotropic properties, and spatial neighborhood information of anatomical structures. Prior to feature extraction, we normalize all MRI intensity ranges to [0, 1], where 1 is assigned to the mean of the highest 5% intensity values of each MR image. The features are explained in detail below.

Location Features. Spatial position of each segmented voxel in three axes, normalized within the given volume, yield 3 location features.

Intensity Features. We extract statistical intensity features from cubic patches with edge length τ_{int} centered at each voxel. We compute intensity mean at two different τ_{int} scales, as well as variance, skewness, and kurtosis at a single τ_{int} scale, yielding a total of 5 intensity features.

Texture and Curvature Maps. Using the image itself (hereafter called the intensity map M_{int}) we extract texture and curvature maps as in [14].

The texture map M_{tex} is generated through convolution of the MRI volume with Gabor filters, which are realized as a combination of a Gaussian filter and a complex sinusoid. The parameters of this texture filter are filter frequency f_{tex}, filter orientation ψ_{tex}, horizontal variance $\sigma_{\text{tex,h}}$, and vertical variance $\sigma_{\text{tex,v}}$. We include M_{tex} as a dimension to our feature space.

For a gray-scale 3D MRI volume, curvature is calculated as the divergence of the normalized intensity gradient. Due to high noise sensitivity of the divergence operation, the intensity map is initially smoothed with a Gaussian kernel with variance σ_{curv}, followed by the removal of gradient values with magnitude smaller than a threshold τ_{curv}. The curvature map M_{curv} is then calculated as the divergence of this smoothed volume.

Anisotropy Features. Anisotropy along an orientation can provide discriminant information. Thus, we compute the anisotropy of both M_{tex} and M_{curv} in the main 7 orientations in 3D (i.e., 4 diagonal and 3 along main axes) using entropy as:

$$\text{ani}^d = -\sum_{\gamma} p_{\gamma}^d \log p_{\gamma}^d \tag{1}$$

where p_{γ}^d is the probability distribution of a texture or curvature value γ within neighborhood $\mathcal{N}_d$ with direction $d \in \{1, .., 7\}$; thus creating 14 features.

Context Features. The presence of one anatomical structure usually allows for the inference of its neighboring structures. In order to capitalize on the regularity of spatial relationships within the human body, we use context features as shifts of intensity, texture and curvature maps in a total of 14 directions (i.e., 8 diagonal and 6 for the main three axes). In order to represent information from a neighborhood of voxels rather than one single point, the image map is initially averaged over a $\mu \times \mu \times \mu$ neighborhood. Context features are extracted from all image maps (M_{int}, M_{tex}, and M_{curv}) at multiple scales of shifts (η_{cont}) around a given center voxel, such that both proximal and distant anatomical relationships can be captured; resulting in a total of ($14\,\text{directions} \times 3\,\text{maps} \times 3\,\text{shifts} =$) 126 context features.

The above lead to a total number of 149 features, a sample subset of which is shown in Fig. 1.

2.2 Sampling

In order to alleviate any bias, an equal number of foreground and background voxels is often sampled for classifier training. However, for full-body MRI volumes, there exist a severe imbalance between the numbers of foreground and background voxels. Conventionally, uniform sampling (UnS) is used, where sample locations are determined with a uniform probability, as illustrated in Fig. 2a. While UnS aims at maximizing classification accuracy, it is not optimal with regard to the final segmentation, as it puts undue emphasis on correctly classifying irrelevant regions further away from the foreground structure; potentially, at the cost of classification accuracy around the anatomical borders, where accuracy is often needed. We therefore choose a sampling approach which is more apt to capturing relevant information needed for the exact delineation of an anatomical structure. We refer to this approach as selective sampling (SeS). In SeS, foreground and background samples are chosen with a probability of

$$\text{p}(x) = \frac{1}{\alpha} \exp\left(-x/\beta_{\text{SeS}}\right) \tag{2}$$

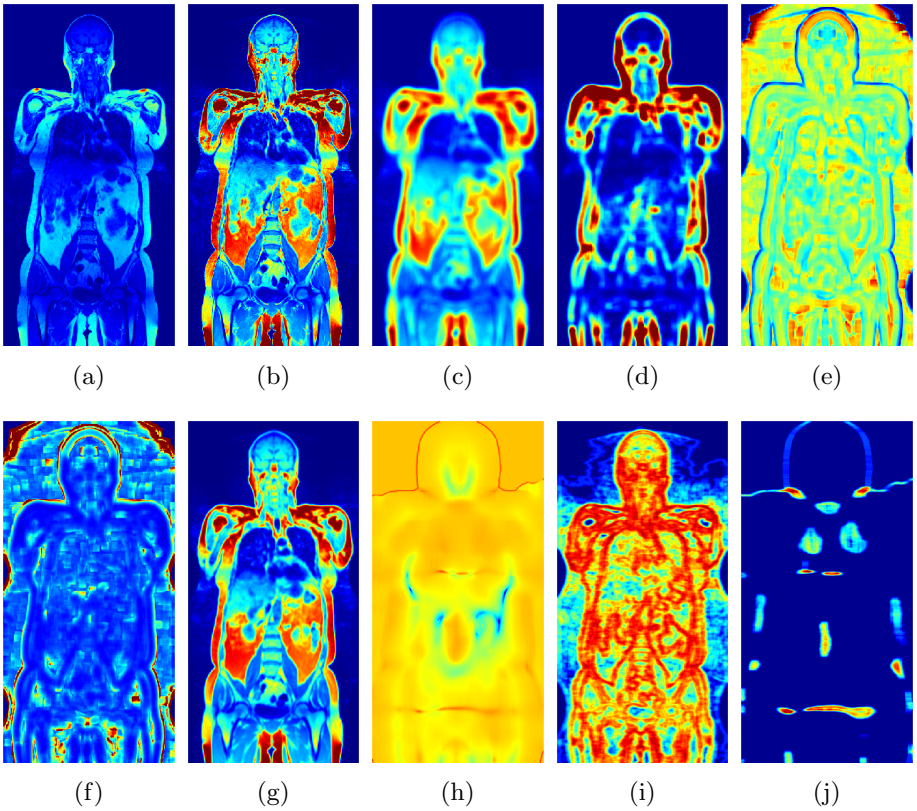
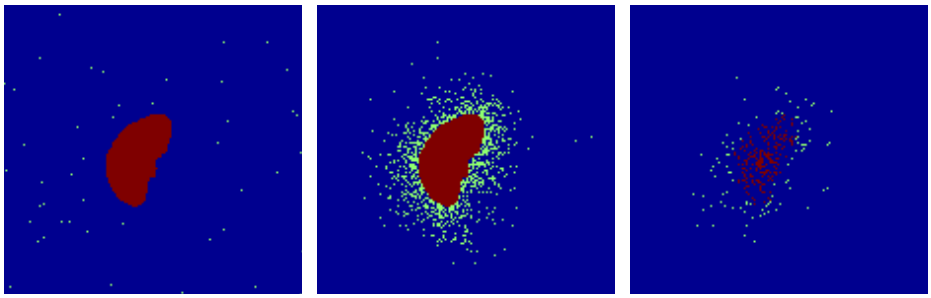


Fig. 1: Example slices of (a) original image M_{int}, (b) mean intensity ($\tau_{int} = 2mm$), (c) mean intensity ($\tau_{int} = 20mm$), (d) intensity variance ($\tau_{int} = 20mm$), (e) intensity skewness ($\tau_{int} = 20mm$), (f) intensity kurtosis ($\tau_{int} = 20mm$), (g) texture map M_{tex}, (h) curvature map M_{curv}, (i) texture anisotropy, and (j) curvature anisotropy.

where α is a normalizing factor, x is the distance to the target structure border and β_{SeS} regulates the exponential decay. Figures 2b and 2c show examples of SeS sample locations for kidney for two different numbers of total samples (n_s).

2.3 Automatic ROI Selection

While learning algorithms can be highly accurate within a local region-of-interest (ROI), with regard to large fields of view, atlas-based segmentation often has the advantage of reliably detecting the approximate position of an organ (leveraging the info from a larger region and hence the entire anatomical appearance). To leverage such approximate localization in our framework, we use a multi-atlas registration framework to transfer atlas annotations to define a ROI for the subsequent classification to work. We use an MRF-based registration approach as



(a) (b) (c)

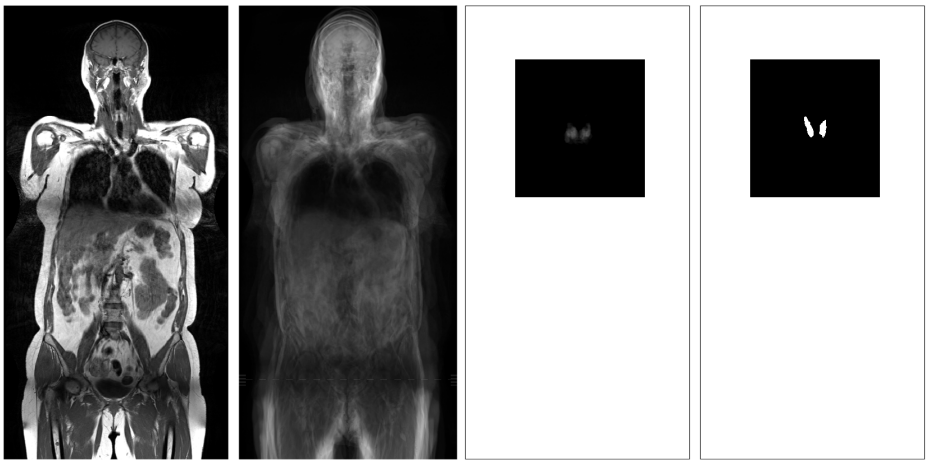
Fig. 2: Cropped example slices of (a) UnS locations with n_s=6000, (b) SeS locations with n_s=6000, and (c) SeS locations with n_s=600. Red pixels indicate sampled foreground locations, while green indicates sampled background. Note that in (a) many green dots are out of the displayed region.

in [8]. For a given testing MRI volume, we register several training MR volumes. A sample superposition of such registered volumes is shown in Fig. 3b, for a test image in Fig. 3a. Accordingly, we also combine the registered annotations for each organ to effectively limit our classification task to a smaller ROI compared to full-body MRI. For a given organ i, the ROI is determined by first finding a bounding box of size $n_x \times n_y \times n_z$ around the registered training annotations, and then enlarging it to a size of $2n_x \times 3n_y \times 2n_z$ determined empirically. Fig. 3c shows the superposition of thyroid training annotations within a constructed ROI, and Fig. 3d shows the ground-truth annotation for this case.

2.4 Random Walker

Mere classification of pixels is likely to create speckled results, with holes and islands of false negatives and positives. These artifacts can be minimized via spatial regularization by incorporating the assumption that neighboring voxels are likely to belong to the same anatomical structure, such as using Markov Random Fields (MRF). Nevertheless, when multiple regions are returned as positives, even regularization cannot determine which one(s) are the desired organ. A common approach is then a post-processing step to select the largest connected component, which however presents no guarantees on the correct selection as the solution of the previous MRF problem is simply taken as a hard-decision (i.e. the probabilities of locations are ignored during the component searching of the final labeling). Random Walker (RW) instead is designed inherently to return a single connected component, which is a an anatomical constraint known in most clinical segmentation tasks. Therefore, in this work we choose RW for spatial regularization.

RW is known to perform successfully, given seed locations inside (FG) and outside (BG) the target structure. Alas, segmentation results can be sensitive to



(a) (b) (c) (d)

Fig. 3: Example slices of (a) the test volume, (b) superposition of registered atlas images, (c) superposition of registered training thyroid annotations within the determined ROI, and (d) test volume ground-truth thyroid annotation.

seed selection. We hereby propose an automatic seed selection method based on mean-shift clustering (MSC) [15]. We initialize n_c clusters centered at voxels with the highest classification scores. Any voxel with a classification score below t_{MSC} is removed from this selection, which leaves $\tilde{n}_c^{(i,s)}$ points for organ i and MRI sequence s. Using the annotated organs from the training set, we calculate the average volume v_{avg} for a given organ. The MSC algorithm is then initiated with $\tilde{n}_c^{(i,s)}$ spheres of volume v_{avg} centered at the selected points. Spheres with mean points closer than half the radius ($r_{(i,s)}/2$) are merged. After MSC convergence, only voxels within the largest cluster are kept and assigned as FG seeds for the RW. BG seeds are placed at a regular interval along the borders of the defined ROI.

3 Results

Data Set. To evaluate the proposed method, the publicly available VISCERAL Anatomy3 data set [12] with 20 full-body T1 sequence MRI images was used. Each of which contains up to 20 annotated anatomical structures. In this paper, we evaluated all structures with six or more available annotations. All results presented in this paper are based on six-fold cross-validation.

Features and Classification. While the method described above does not call for a specific type of classifier and may easily be extendable to other learning methods, we have used RF classifiers for all experiments presented in this paper.

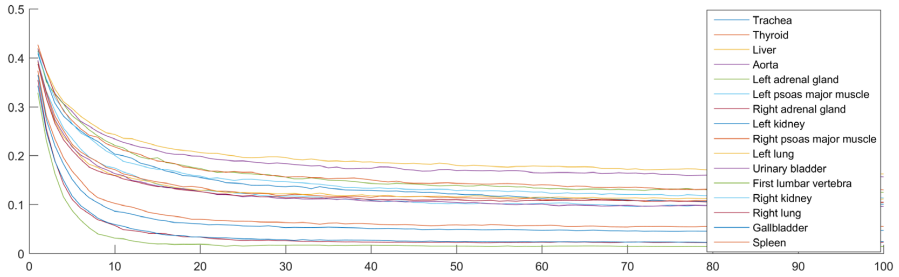


Fig. 4: Out-of-bag error vs. forest size for random forests trained on MRI of anatomical structures.

This choice was based on preliminary experiments with RF and linear support vector machines. We have trained separate RF classifiers for each organ, with feature and classifier parameters determined empirically with a set of preliminary experiments on smaller image/organ sets. RF tree size was set to 50, as derived from the out-of-bag error shown in Fig. 4. Similarly, 3000 foreground and 3000 background samples were used for training each RF classifier. Different types of features were analyzed using RF variable-importance scores. We set the feature extraction parameters as follows: *Intensity features:* Mean with neighborhood size τ_{int}=2 mm and τ_{int}=20 mm, variance, skewness, and kurtosis with neighborhood size τ_{int}=20 mm. *Texture map:* Horizontal variance $\sigma_{\text{tex,h}}$=2, vertical variance $\sigma_{\text{tex,v}}$=4, frequency f_{tex}=16, and orientation $\psi_{\text{tex}} = \frac{\pi}{2}$. *Curvature map:* Variance σ_{curv}=16 and threshold τ_{curv}=0.01. *Anisotropy features:* Neighborhood size η_{ani}=20 mm. *Context features:* Averaging neighborhood size μ_{cont}=3 mm, extraction scales η_{cont}=4 mm, η_{cont}=8 mm and η_{cont}=16 mm.

Registration for ROI Selection. For multi-atlas ROI selection, an MRF-based registration framework as in [8] was used, which yields successful ROI definitions for larger organs. For some organs (in particular, gallbladder and adrenal glands), for which shape and anatomical surrounding are not good indicators and for which such atlas-based segmentations typically fail, we switched to affine transformation of the entire anatomy for the multi-atlas approach for a less precise but robust location prior.

Sampling and Spatial Regularization. In order to analyze the influence of sampling and spatial regularization on segmentation performance, we performed experiments with the proposed SeS sampling and RW spatial regularization as well as the corresponding standard approaches: UnS sampling as well as MRF spatial regularization. The SeS sampling probability decay factor β_{SeS}=12 mm was optimized in preliminary experiments using a subset of anatomical structures. For MRF evaluation, RF classifier scores were used as MRF unary costs and the pairwise costs were constructed by Potts model from MR intensity images. The proposed RW spatial regularization was parameterized empirically as n_c=10000 and t_{MSC}=0.92.

Table 1: Organ Dice scores before spatial regularization and after RW/MRF when trained with UnS and SeS vs. inter annotator median (inA_{med}) from [12].

		Adrenal gland-L	Adrenal gland-R	Aorta	Lumbar Vert. 1	Gallbladder	Kidney-L	Kidney-R	Liver	Lung-L	Lung-R	Psoas major-L	Psoas major-R	Spleen	Thyroid	Trachea	Urinary Bladder	
UnS	RF	0.05	0.10	0.12	0.06	**0.05**	0.17	0.14	0.51	0.49	0.62	0.15	0.13	0.21	0.05	0.09	0.21	soa
	MRF	0.05	0.08	0.13	0.06	**0.05**	**0.16**	0.14	**0.57**	0.56	0.68	0.16	0.14	0.23	0.09	0.13	0.26	
	RW	**0.09**	**0.17**	**0.16**	**0.10**	**0.05**	**0.16**	**0.17**	0.44	**0.63**	**0.69**	**0.26**	**0.21**	**0.28**	**0.12**	**0.20**	**0.37**	Proposed
SeS	RF	0.61	0.50	0.06	0.09	0.63	0.12	0.08	0.30	0.19	0.25	0.11	0.12	0.12	0.22	0.07	0.03	
	MRF	0.26	0.29	**0.44**	0.12	**0.68**	0.44	0.35	0.62	0.41	0.49	0.54	0.51	0.48	0.35	0.48	0.60	
	RW	**0.64**	**0.75**	0.39	**0.14**	0.60	**0.58**	**0.59**	**0.76**	**0.76**	**0.79**	**0.66**	**0.63**	**0.68**	**0.45**	**0.50**	**0.73**	
inA_{med}		0.61	0.55	0.82	0.76	0.74	0.91	0.91	0.90	0.94	0.93	0.85	0.85	0.75	-	0.78	0.90	

Table 2: Relative Dice improvement of the proposed SeS-RW over other methods for each anatomical structure given in the same order as Table 1.

SeS-RW/UnS-MRF	12.80	9.38	3.00	2.33	12.00	3.62	4.21	1.33	1.36	1.16	4.12	4.50	2.96	5.00	3.85	2.81
SeS-RW/UnS-RW	7.11	4.41	2.44	1.40	12.00	3.62	3.47	1.73	1.21	1.14	2.54	3.00	2.43	3.75	2.50	1.97
SeS-RW/SeS-RF	1.05	1.50	6.50	1.56	0.95	4.83	7.37	2.53	4.00	3.16	6.00	5.25	5.67	2.05	7.14	24.33
SeS-RW/SeS-MRF	2.46	2.59	0.89	1.17	0.88	1.32	1.69	1.23	1.85	1.61	1.22	1.24	1.42	1.29	1.04	1.22

The average of 6-fold cross-validation Dice scores per organ for all four combinations UnS-MRF, UnS-RW, SeS-MRF, and SeS-RW as well as the classification F1 scores (rows labeled as RF) prior to spatial regularization for both UnS and SeS are presented in Table 1 along with the inter annotator median Dice score reported in Fig. 4 of [12]. The highest scores per anatomical structure for UnS and SeS separately are indicated in bold, while the highest score for both is highlighted in gray. Methods presented in this paper as novel combinations for MRI segmentation are shaded in light blue. Out of these, we propose in particular SeS-RW, selective sampling with random-walker. Accordingly, in Table 2 we present the relative Dice score comparison of SeS-RW over other method combinations, where UnS-MRF is taken as the state-of-the-art of similar classification-based methods not utilizing the techniques proposed herein.

4 Discussion

Quantitative results presented in Table 1 show that, for the given data set, RW with the proposed automatic seed selection outperforms the commonly used MRF regularization. Furthermore, comparative results in Table 2 indicate that the proposed approach provides on average over 12× Dice improvement for some smaller organs such as gallbladder and adrenal gland, with a minimum average improvement being 16% for the right lung. Prior to spatial regularization, UnS outperforms SeS for 11 out of 16 anatomical structures. This is due to the

boundary-focus of SeS, which leads to a large number of false positives outside the organ to segment. However, after spatial regularization, methods based on SeS prevail for all anatomical structures. It can be seen that the best results are achieved by combining SeS sampling with RW. These scores indicate that the false positives outside the target anatomical structure, caused by the focus of the classifier on organ border regions, were successfully recovered by the proposed spatial regularization strategy.

As seen in Table 1, for some anatomical structures, such as the aorta or the first lumbar vertebra, our resulting Dice scores are relatively low. We attribute this to the large amount of contextual information necessary to successfully segment these structures. Clinical protocols for aorta segmentation, for example, define the extents of the organ in relation to other structures. In such cases an atlas-based segmentation as expected performs better through leveraging the mutual contextual information from other anatomical structures.

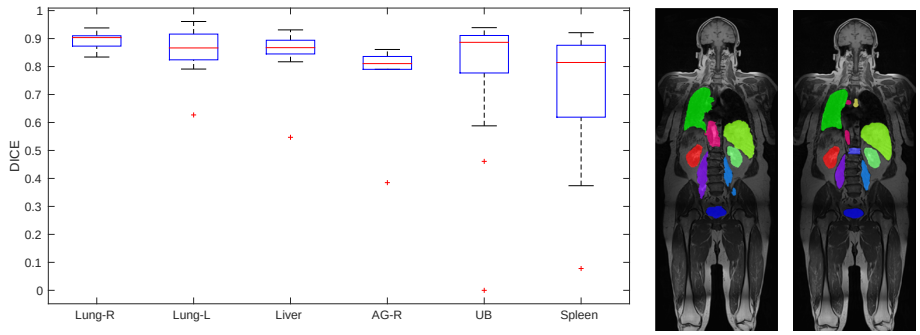
While there currently is no atlas-based segmentation baseline for the VISCERAL Anatomy3 full-body MR T1 modality, [3] lists such scores for the predecessor challenges Anatomy1 and Anatomy2. Comparing the results in [3] to SeS-RW confirms that atlas-based segmentation is superior for organs such as first lumbar vertebra or aorta. On the other hand, however, SeS-RW is more promising for smaller or irregularly shaped organs such as both adrenal glands, gallbladder, spleen and urinary bladder. Atlas-based segmentation as used in [3] tends to achieve higher scores for the right lung as opposed to the left, which we assume to be caused by the larger portion of heart located on a person's left side. In contrast, SeS-RW's scores for left and right lung are almost identical.

In the future, we will integrate atlas-segmentation priors, which we now already use for the ROI selection, as a probabilistic location prior into our SeS-RW framework in order to combine the benefits of both frameworks. Nevertheless, our current SeS-RW approach stands out for high performance in segmenting some difficult organs, e.g. adrenal glands and gallbladder, even in comparison to CT segmentations. For the adrenal glands, we are at a level of accuracy comparable to inter annotator variability.

In Fig. 5a, we show the Dice score distribution of some representative anatomical structures with our SeS-RW method. Additionally, a sample segmentation result of our SeS-RW algorithm on one case is shown in Fig. 5b for qualitative comparison with the corresponding ground-truth annotations shown in Fig. 5c for this particular plane.

5 Conclusions

In this paper, we presented an automatic segmentation framework for abdominal organs. We demonstrated the significance of choosing our sampling method in accordance with the proposed spatial regularization. While all our experiments were performed on MRI images, due to the complexity of MRI, we believe that this approach can be generalized for other imaging modalities and anatomical structures (i.e. bones) which were not part of our experiment data. Our future



(a) (b) (c)

Fig. 5: (a) Distribution of Dice scores with SeS-RW segmentation, where AG is adrenal gland and UB is urinary bladder. (b) A sample SeS-RW segmentation and (c) its ground-truth annotation.

work will focus on combining atlas-based registration (i.e. from [8]) with our framework in a probabilistic manner to further improve segmentation accuracy.

6 Acknowledgements

This work was funded by the Swiss National Science Foundation (SNSF) and the Highly Specialized Medicine (HSM) project of Zurich Department of Health.

References

1. Kahl, F., Alvén, J., Enqvist, O., Fejne, F., et al.: Good features for reliable registration in multi-atlas segmentation. In: ISBI, 12–17 (2015)
2. Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D.: Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE Trans Med Imaging 32, 1723–1730 (2013)
3. Gass, T., Szekely, G., Goksel, O.: MICCAI, chapter Multi-atlas Segmentation and Landmark Localization in Images with Large Field of View, 171–180. Springer International Publishing (2014)
4. Heinrich, M.P., Maier, O., Handels, H.: Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. In: ISBI, 27–30 (2015)
5. He, B., Huang, C., Jia, F.: Fully automatic multi-organ segmentation based on multi-boost learning and statistical shape model search. In: ISBI, 18–21 (2015)
6. Okada, T., Linguraru, M.G., Hori, M., Suzuki, Y., et al.: Multi-organ segmentation in abdominal ct images. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 3986–3989. IEEE (2012)
7. Chen, X., Udupa, J.K., Bagci, U., Zhuge, Y., Yao, J.: Medical image segmentation by combining graph cuts and oriented active appearance models. IEEE Transactions on Image Processing 21, 2035–2046 (2012)

8. Gass, T., Szkely, G., Goksel, O.: Simultaneous segmentation and multiresolution nonrigid atlas registration. IEEE Transactions on Image Processing 23, 2931–2943 (2014)
9. Luo, Q., Qin, W., Wen, T., Gu, J., et al.: Segmentation of abdomen mr images using kernel graph cuts with shape priors. Biomed Eng Online 12, 1–19 (2013)
10. Yang, X., Minh, H., Cheng, T., Sung, K.H., Liu, W.: MICCAI, chapter Automatic Segmentation of Renal Compartments in DCE-MRI Images, 3–11. Springer International Publishing (2015)
11. Mukhopadhyay, A., Oksuz, I., Bevilacqua, M., Dharmakumar, R., Tsaftaris, S.A.: MICCAI, chapter Unsupervised Myocardial Segmentation for Cardiac MRI, 12–20. Springer International Publishing (2015)
12. Jimenez-del-Toro, O., Muller, H., Krenn, M., Gruenberg, K., et al.: Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. IEEE Transactions on Medical Imaging PP, 1–1 (2016)
13. Breiman, L.: Random forests. Machine learning 45, 5–32 (2001)
14. Mahapatra, D., Schuffler, P.J., Tielbeek, J.A.W., Makanyanga, J.C., et al.: Automatic detection and segmentation of crohn's disease tissues from abdominal MRI. IEEE Trans Med Imaging 32, 2332–2347 (2013)
15. Fukunaga, K., Hostetler, L.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inf Theory 21, 32–40 (1975)