Hierarchical multi-organ segmentation without registration in 3D abdominal CT images

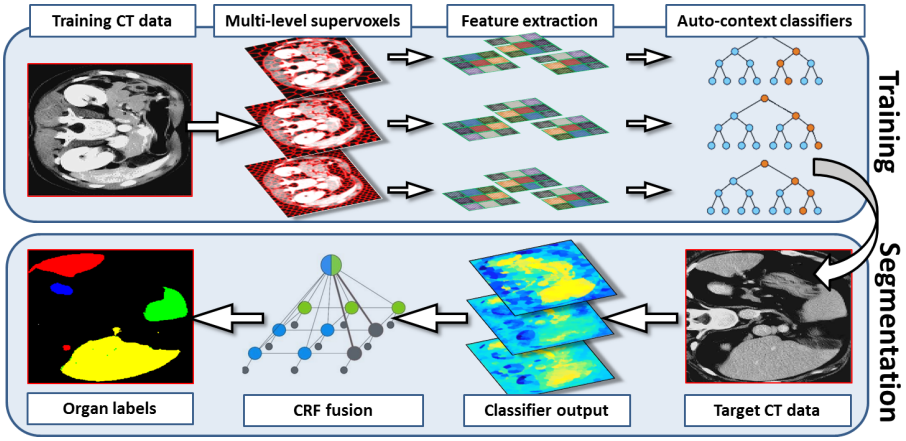
Vasileios Zografos[1], Alexander Valentinitsch[1,2], Markus Rempfler[1], Federico Tombari, and Bjoern Menze[1]

[1] Computer Aided Medical Procedures & Augmented Reality, TUM, Germany
[2] Department of Diagnostic and Interventional Neuroradiology, TUM, Germany

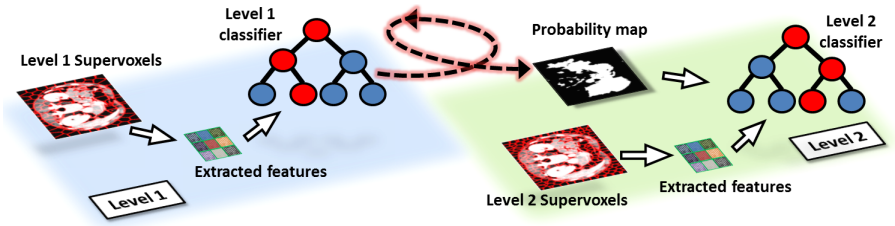
Abstract. We present a novel framework for the segmentation of multiple organs in 3D abdominal CT images, which does not require registration with an atlas. Instead we use discriminative classifiers that have been trained on an array of 3D volumetric features and implicitly model the appearance of the organs of interest. We fully leverage all the available data and extract the features from inside supervoxels at multiple levels of detail. Parallel to this, we employ a hierarchical auto-context classification scheme, where the trained classifier at each level is applied back onto the image to provide additional features for the next level. The final segmentation is obtained using a hierarchical conditional random field fusion step. We have tested our approach on 20 contrast enhanced CT images of 8 organs from the VISCERAL dataset and obtained results comparable to the state-of-the-art methods that require very costly registration steps and a much larger corpus of training data. Our method is accurate, fast and general enough that may be applied to a variety of realistic clinical applications and any number of organs.

1 Introduction

Multiple organ segmentation in abdominal computer tomography (CT) images can be an important step to computer aided diagnosis and computer assisted surgery. Existing work in automated multi-organ segmentation can be roughly divided into *registration based* and *classification based*. Registration methods include statistical shape models (SSM) [13], probabilistic atlases (PA) [4, 12] and multi-altas techniques (MA) [16]. SSM approaches work by employing several shape or appearance models, usually in conjunction with hierarchical object localisation. Although SSM can produce accurate segmentations, they require good initialisation otherwise registration between the SSM the organs will fail. PA are more robust to registration with a target image since they incorporate global spatial information as well as inter-organ spatial relationships. However, both SSM and PA cannot handle large inter-subject variabilities, so research has moved on to target-specific MA solutions, which have shown to be superior to single model/atlas approaches. All registration-based methods are limited in



(a) The main components in our registration-free CT segmentation approach.



(b) Hierarchical auto-context between two subsequent levels.

Fig. 1. Proposed method outline (a) and details of the hierarchical auto-context (b)

that they require every organ to be present and to have stable locations between training and test images (localised spatial support). Furthermore, non-rigid registration can be very time consuming. Especially for MA approaches, it is necessary to have all the atlases available during segmentation time and register the target image with each atlas separately. The classification-based methods on the other hand, are not plagued by the same problems as registration-based approaches. Instead, they can predict the probability that a voxel belongs to a specific object based on previously seen data. Most classification-based methods [11, 5] use some flavour of the random forest classifier and are trained by local appearance features. Furthermore, issues such as non-localised spatial support and large inter-subject variability may be dealt with by training with additional data. Even though classification-based methods can be fast, they do not take into consideration organ contextual information or organ shape and as such they often produce less accurate segmentations than registration-based methods.

This paper proposes a novel framework for multi-organ segmentation (Fig. 1(a)), which leverages several ideas from computer vision and machine learning and does not require any registration steps, neither during training nor during

segmentation time. Because of this, we can avoid all the potential shortcomings of registration-based methods, while at the same time design a method that is accurate and fast enough that can be applied to real-life clinical applications. We begin by generating supervoxels from the CT image at multiple levels of detail (Sec. 2.1). Then we extract a set of complementary appearance and contextual features from the supervoxels (Sec. 2.2) and use them to train a boosted tree classifier at each level. The classifiers are not independent but are linked together using hierarchical auto-context (Sec. 2.3). During segmentation, the linked classifiers are applied to the new image and their output is fused using a hierarchical conditional random field (Sec. 2.4). We have tested our approach for the segmentation of 8 organs in a 20 CT image dataset (Sec. 3) and obtained results comparable to the state-of-the-art registration-based methods, despite our solution being registration-free. Also, we are considerably more efficient than most competitors since training is done offline and the training data does not have to be present during segmentation. Our key contributions are:

- A method for **registration-free** multi-organ segmentation in 3D CT images
- Multiple levels of **supervoxels** for appearance and context learning
- Adaptation of **auto-context** to a hierarchical scenario
- Extension of **3D feature descriptors** to volumetric data
- **CRF fusion** using spatial and hierarchical supervoxel neighbourhoods

2 Method

2.1 Multi-level supervoxels for learning appearance and context

The first step, after acquiring and pre-processing the training data, is to generate a supervoxel representation within which we may extract the appearance features. A supervoxel representation is simply an oversegmentation of the image into homogeneous regions and it is carried out by grouping adjacent voxels based on their intensity similarities. Given a 3D image with voxels $\mathbf{v}=1:V$, we can define a supervoxel as the set of voxels $\mathcal{S}_l = \{\mathbf{v} : s(\mathbf{v}) = l\}$, where $l = 1:L$ and $s : \{1, ..., V\} \rightarrow \{1, ..., L\}$. We have used the fast implementation by [7]. Working with supervoxels is preferable to using single voxels or arbitrary patches, since: 1) we have better adherence to object boundaries and as such are more likely to preserve these boundaries in the final segmentation; 2) The homogeneous regions inside each supervoxel usually come from a single organ since the shape and size of a supervoxel adapts to the local information. Extracting therefore features from inside each supervoxel means that we can capture the specific local structures of individual organs, from a more natural voxel neighbourhood and without confounding information from different organs; Finally, 3) using supervoxels instead of voxels means that we have a reduced model complexity, which in turn results in a much faster algorithm. Instead of using only a single grid to generate the supervoxels, we have adopted a multi-level approach whereby we apply multiple initialisation grids (in a coarse-to-fine strategy), in order to obtain supervoxels at various sizes, shapes and granularity. Our aim with this

multi-level approach is to capture a more diverse set of local structures at multiple scales and from different-sized neighbourhoods, in order to obtain a richer representation of organs that may exhibit a large variation in appearance.

2.2 3D volumetric feature extraction

Unequivocally, the most important part of our framework is the choice of features used to train the classifiers, since they directly influence the accuracy of the segmentation. We have extracted a mixture of texture, shape and neighbourhood context features in order to obtain a comprehensive representation of the organs of interest. We denote a feature vector as: $\mathbf{d} = \{\mathbf{d}^{\mathrm{G}}, \mathbf{d}^{\mathrm{V}}, \mathbf{d}^{\mathrm{H}}, \mathbf{d}^{\mathcal{N}}\}$.

3D GLCM features: The gray level co-occurrence matrix (GLCM) is a commonly used approach [10] for extracting statistical features between pixels/voxels in image data. We have adapted this idea to supervoxels, where each entry in the 3D GLCM represents the probability of different graylevels occurring between neighbouring voxels. The neighbourhoods are defined inside a supervoxel and the displacement between two voxels is given by the vector $\{d, \theta, \varphi\}$, where d is the $\mathcal{L}_1$ distance and $\{\theta, \varphi\}$ are the azimuth and zenith angles that determine direction in 3D polar coordinates. For simplicity, we set d=1 and calculated 13 combinations of corresponding directions. This gives 13 Harralick-type features and for each such feature we have extracted both the angular mean and standard deviation, resulting in a 26-dimensional texture vector for every supervoxel.

Volumetric Shape Context features: The 3D shape context (3DSC) feature [6] is a histogram that accumulates the number of shape points within a given volume. We have extended the 3DSC, originally proposed for 3D point clouds and meshes, to work with volumetric data. We denote this as the Volumetric Shape Context (VSC) descriptor. The VSC uses a 3D gradient intensity histogram centred around each voxel. However unlike the 3DSC, the histogram is now a cube regularly subdivided along its three dimensions, so that each bin describes the same portion of 3D space and contains the same number of voxels. The volume of the cube is given by the volume of the associated supervoxel inside which the current voxel resides. In addition, we assume a global 3D coordinate frame that remains consistent amongst the acquired data. Given thus the gradient $\nabla f(\mathbf{v})$ of a voxel $\mathbf{v}$ at coordinates (v_x, v_y, v_z), each bin $h(k)$ of the histogram stores the average gradient computed from all the N voxels falling within the associated volume of the cube $C(k)$:

$$h(k) = \frac{1}{N} \sum_{\mathbf{v} \in C(k)} \nabla f((v_x, v_y, v_z)). \tag{1}$$

HOG3D features: The HOG3D is a local descriptor based on oriented histograms of 3D gradients and is complementary to the VSC. We have used the algorithm by [8] and have adapted it to volumetric data and supervoxels. The computation of the descriptor involves first calculating the 3D gradients around a point of interest. Then, the orientation of these gradients is quantised using regular polyhedra and the mean gradient is computed. In our case, the point

of interest is the supervoxel centroid. The gradients are computed and averaged over the spatial support of the supervoxel. HOG3D features differ from VSC in that the former uses multiple histograms and accumulates gradient orientations, while the latter has a single histogram and accumulates gradient intensities.
Neighbourhood context: One simple way of including additional discriminative power into the algorithm is to relate nearby supervoxels together, thereby incorporating neighbourhood context information. This is because in general, the organs of interest have stable relative positions and so we also expect that relative contextual information between supervoxels to be consistent between training and test images. We may define a neighbourhood $\mathcal{N}$ around a given supervoxel $\mathcal{S}_l$ as those supervoxels that share a boundary with $\mathcal{S}_l$. Then for every supervoxel $\mathcal{S}_n$, $n \in \mathcal{N}$ inside the neighbourhood, we calculate the difference $\mathbf{D}_n = \|\mathbf{d}_l - \mathbf{d}_n\|_1$ at each feature-type. Since the size of the neighbourhood can vary for different supervoxels, we only consider the mean and the maximum of $\mathbf{D}_n$, giving us two neighbourhood context features for each supervoxel. Therefore, for every supervoxel we extract a 177-dim vector $\mathbf{d}$, which is composed by concatenating the 26-dim GLCM features $\mathbf{d}^{\mathrm{G}}$ the 125-dim VSC features $\mathbf{d}^{\mathrm{V}}$ the 20-dim HOG3D features $\mathbf{d}^{\mathrm{H}}$ and the 6-dim neighbourhood context features $\mathbf{d}^{\mathcal{N}}$.

2.3 Hierarchical auto-context classification

After feature extraction the next step is to train the classifiers. Here we use the gradient boosted trees (GBT) [14], which is an ensemble prediction method where boosting is applied to weak decision trees. GBTs can often surpass generic random forests and produce a very good fit to the data even in the case of complex nonlinear problems. In order to incorporate all the information contained in the features from the different supervoxels levels, it makes more sense to link the levels together than to treat each level independently. We therefore train one GBT classifier for each supervoxel level and link them using a technique called *auto-context* [15]. In auto-context a classifier is first trained from local features and then applied back onto the image to produce discriminative probability maps. These maps, which act as a rough object localiser, are appended to the existing local features and are used to train a new classifier.

We have introduced two novelties to the basic auto-context algorithm. First, we have extended it to work in a hierarchical scenario and thereby linking together the classifiers from all the levels. More specifically we train the initial GBT from the features extracted at the coarsest level and apply it back onto the image to produce a probability map. The probability map is then transferred to the image on the next level (from coarse supervoxels to fine supervoxels) and together with the features extracted at this new level are used to train a new GBT classifier (Fig. 1(b)). This procedure is repeated until we reach the final level. The output of the training stage is a set of linked GBT classifiers and the output of the segmentation stage is a set of probability maps. The probability maps will be merged in the final CRF fusion step. The second extension is that we further exploit the information in the probability maps and use them as importance sampling weights. Namely, at each subsequent level we only train with

the features at locations where the preceding classifier had a high confidence. This step avoids inundating the classifier with too much data and additionally increases the confidence of the classifier at every level by only training with strong, discriminative features.

2.4 Hierarchical CRF fusion using supervoxel neighbourhoods

The last component of our framework is a conditional random field (CRF) step where the hierarchical outputs from the auto-context classification are fused in order to determine the best labelling. CRF fusion is by far superior to other merging approaches such as voting or averaging. The CRF structure is specified by the undirected graph $\mathcal{G}=(\mathcal{V}, \mathcal{E}_a \cup \mathcal{E}_p)$, where $\mathcal{E}_a=\{(i,j) \in \mathcal{S} \times \mathcal{S} |\ i$ is **adjacent** to $j\}$ and $\mathcal{E}_p=\{(i,j) \in \mathcal{S} \times \mathcal{S} |\ i$ is **parent** to $j\}$. $\mathcal{E}_a$ contains all pairs of supervoxels that are neighbours on the same level, whereas $\mathcal{E}_p$ is the set of all parent-child supervoxel pairs between two subsequent levels. The energy function is given by:

$$E(\mathbf{y}) = \sum_{i \in S} \phi_i(y_i) + \sum_{(i,j) \in \mathcal{E}_a} \phi^a_{i,j}(y_i, y_j) + \sum_{(i,j) \in \mathcal{E}_p} \phi^p_{i,j}(y_i, y_j), \tag{2}$$

where y are the labels. Hence, the CRF introduces both spatial regularisation *within* each level of supervoxels by ϕ^a as well as interaction potentials *between* levels by ϕ^p. We use the probabilistic output $P(y_i|\mathbf{d}_i)$ of the classifiers for the unary potentials:

$$\phi_i(y_i) = -\log P(y_i|\mathbf{d}_i), \tag{3}$$

where $\mathbf{d}$ are the extracted feature vectors we have used to train the classifiers with. The binary potentials are set as:

$$\phi_{i,j}(y_i, y_j; \lambda) = \lambda \exp(-\gamma ||\mathbf{d}_i - \mathbf{d}_j||_1)(1 - \delta(y_i, y_j)), \tag{4}$$

with γ=1/ dim($\mathbf{d}$) and $\delta(.)$ being the Kronecker delta function. λ is a scalar parameter that is chosen separately for the spatial and the hierarchical potentials. We estimate the best labelling $\mathbf{y}^\star$ by minimising (2) with the algorithm of [9].

3 Experiments and results

Dataset: We have used the VISCERAL Anatomy3 dataset [1], which includes 20 contrast enhanced, abdominal CT images. Each CT image has a resolution of 512×512 pixels with an average of 426 slices and a resolution between 0.604-0.793 mm. The images are manually segmented and the ground truth annotations contain up to 20 anatomical structures, albeit not ubiquitous. We will consider 8 organs here: liver, spleen, 2 kidneys, pancreas, 2 lungs, urinary bladder; because they are the most consistently represented in the dataset. In order to improve the appearance learning and discriminative power of the classifier we utilised a secondary add-on dataset, the VISCERAL Silver Corpus [2]. This dataset contains an additional 59 useful CT images, but without any manual ground

truth. Instead the labels have been automatically obtained and as such contain segmentation errors. Despite that, the data can still be used for noisy training since the errors are mostly manifested as organ under-segmentations. This means that if we ignore the background information we may still incorporate the partial organ labels from the Silver Corpus.

Pre-processing: Every image was first downsized by a factor of 2 and cropped to speed up training and segmentation. Following that, we converted the data to an isotropic resolution, windowed the Houndsfield units between [0,150] and mapped to intensities in the range [0,1]. Finally, we performed histogram equalisation and denoised the images using 3D anisotropic diffusion.

Training and segmentation: We defined 9 classes, one for each of the 8 organs and a background class for all the remaining structures. Features were extracted at 4 different levels of detail with 5k, 10k, 20k and 30k supervoxels respectively. The classifiers were set to run for 300 iterations using an exponential loss function and a tree depth of 2. We followed a leave-one-out evaluation strategy, in which the classifiers were trained on 19(+59 noisy) examples and tested on 1. The final organ labels were obtained by the CRF fusion with fixed parameters λ=0.05 for both the hierarchical and spatial potentials.

Results: The main results from our experiments are presented in Table 1 with exemplar segmentation in Fig. 2. We see that our approach obtains good segmentation for the majority of the organs. Furthermore, our results are on par with other state-of-the-art methods from literature that use much larger datasets. Although we cannot yet fully outperform the very accurate registration-based methods [16, 4, 17, 12, 13] we expect that our results will improve upon increasing the noise-free training data to comparable sizes. Note however that we are considerably better than the classification-based method [11] that does not leverage additional information from the data like we do. For reference, we have also included (last column, Table 1) the average results from the methods participated in the VISCERAL Anatomy2 segmentation challenge [3]. This dataset is closely related to ours and so direct comparison is more reasonable. We observe that our method compares very favourably to the average results reported in [3].

4 Conclusions

We have presented a novel classifier-based framework for registration-free multi-organ segmentation in CT images. We have adapted and extended a number of concepts such as multi-level supervoxels, hierarchical auto-context and CRF fusion, in order to fully leverage all the available information and improve the segmentation quality. Our method was evaluated on a 20 image contrast enhanced CT dataset for the segmentation of 8 organs. In terms of accuracy our results are comparable with other state-of-the-art methods that use a much larger corpus of training data. Also, because our training is done offline and

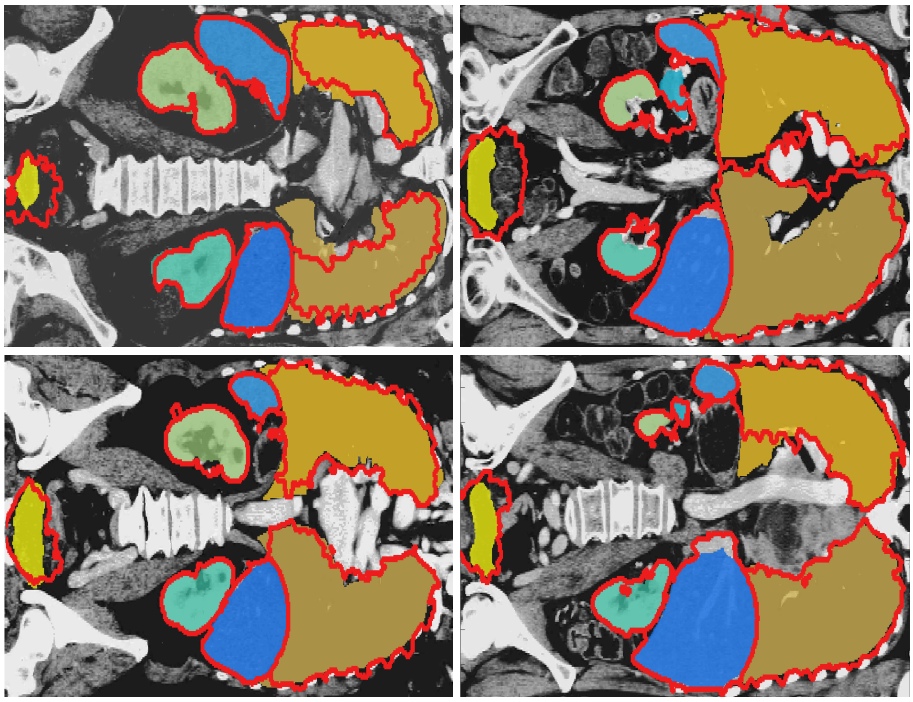


Fig. 2. From left to right: (upper row) Worst, bottom 10%; (lower row) top 10% and best results from the 20 evaluated CT images. Our segmentations are outlined in red over the manual labels.

is decoupled from the segmentation stage, we can increase accuracy by training with more data but without any additional segmentation cost. This is not possible for registration-based methods because they do not scale very well with increasing data. Moreover other approaches require all the atlases to be available during segmentation time. All we need to carry over is a small set of trained classifiers with a minimal memory footprint and without data storage and privacy issues. This makes our method efficient, portable and very practical.

References

1. http://www.visceral.eu/benchmarks/anatomy3/
2. http://www.visceral.eu/assets/Uploads/Deliverables/VISCERAL-D-3-3.pdf
3. Proc. of the VISCERAL benchmark. In: IEEE ISBI (2014)
4. Chu, C., Oda, M., et al.: Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images. In: MICCAI (2013)
5. Cuingnet, R., Prevost, R., et al.: Automatic Detection and Segmentation of Kidneys in 3D CT Images Using Random Forests. In: MICCAI (2012)
6. Frome, A., Huber, D., Kolluri, R., Bulow, T., Malik, J.: Recognizing objects in range data using regional point descriptors. In: ECCV. vol. 3 (2004)

Method	Wang [16]	Chu [4]	Wolz [17]	Oda [12]	Okada [13]	Lombaert [11]	Our	[3]
CT type	?	CTce	CTce	?	CTce	mixed	CTce	CTce
Data size	**100**	**100**	**150**	**100**	**28**	**250**	**20**	**5**
Liver	89.57	90.6	88.9	89.0	89.1	73.2	83.68	83.77
R. kidney	85.87	82.3	86.8	80.8	88.2	28.1	86.74	77.30
L. kidney					87.4	29.4	85.37	80.10
Pancreas	48.69	54.6	55.5	42.1	46.6	-	42.30	23.90
Spleen	86.04	84.5	86.2	74.5	82.5	44.5	84.84	65.99
R. lung	-	-	-	-	-	88.4	81.17	92.51
L. lung						85.3	78.20	92.31
Bladder	-	-	-	-	-	-	59.77	60.77

Table 1. Jaccard indices of different multi-organ segmentation methods. The numbers have been obtained from their respective publications.

7. Holzer, M., Donner, R.: Over-segmentation of 3D medical image volumes based on monogenic cues. In: CVWW. pp. 35–42 (2014)
8. Kläser, A., Marszaek, M., Schmid, C.: A spatio temporal descriptor based on 3D Gradients. In: BMVC (2008)
9. Komodakis, N., et al.: Performance vs computational efficiency for optimizing single and dynamic MRFs: Setting the state of the art... CVIU 112(1), 14–29 (2008)
10. Kovalev, V.A., Kruggel, F., Gertz, H.J., von Cramon, D.Y.: Three-dimensional texture analysis of MRI brain datasets. IEEE TMI 20(5), 424–433 (2001)
11. Lombaert, H., Zikic, D., Ayache, A.C.N.: Laplacian forests: Semantic image segmentation by guided bagging. In: MICCAI. vol. 8674, pp. 496–504 (2014)
12. Oda, M., et al.: Organ segmentation from 3D abdominal CT images based on atlas selection and graph cut. In: Abdominal Imaging (2012)
13. Okada, T., et al.: Abdominal multi-organ segmentation of CT images based on hierarchical spatial modeling of organ interrelations. In: Abdominal Imaging. CCA. vol. 7029, pp. 173–180 (2012)
14. Sznitman, R., Becker, C., Fleuret, F., Fua, P.: Fast object detection with entropy-driven evaluation. In: CVPR (2013)
15. Tu, Z.: Auto-context and its application to high-level vision tasks. In: CVPR (2008)
16. Wang, Z., Bhatia, K., Glocker, B., Marvao, A., Dawes, T., Misawa, K., Mori, K., Rueckert, D.: Geodesic patch-based segmentation. In: MICCAI (2014)
17. Wolz, R., Chu, C., Misawa, K., Fujiwara, M.: Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE TMI 32(9) (2013)