Multi-Atlas-Based Segmentation With Local Decision Fusion—Application to Cardiac and Aortic Segmentation in CT Scans

Ivana Išgum*, Marius Staring, Annemarieke Rutten, Mathias Prokop, Max A. Viergever, and Bram van Ginneken

***Abstract*—A novel atlas-based segmentation approach based on the combination of multiple registrations is presented. Multiple atlases are registered to a target image. To obtain a segmentation of the target, labels of the atlas images are propagated to it. The propagated labels are combined by spatially varying decision fusion weights. These weights are derived from local assessment of the registration success. Furthermore, an atlas selection procedure is proposed that is equivalent to sequential forward selection from statistical pattern recognition theory. The proposed method is compared to three existing atlas-based segmentation approaches, namely 1) single atlas-based segmentation, 2) average-shape atlas-based segmentation, and 3) multi-atlas-based segmentation with averaging as decision fusion. These methods were tested on the segmentation of the heart and the aorta in computed tomography scans of the thorax. The results show that the proposed method outperforms other methods and yields results very close to those of an independent human observer. Moreover, the additional atlas selection step led to a faster segmentation at a comparable performance.**

***Index Terms*—Aortic segmentation, atlas-based segmentation, cardiac segmentation, registration, segmentation by registration.**

I. Introduction

RELIABLE quantitative analysis in medical images, e.g., volume measurements, require delineation of anatomical structures. This is a difficult task, often performed by a human operator. Manual segmentation is time consuming, thus difficult to perform in very large number of scans, for example, obtained in screening programs. Due to technological developments the number of images to be analyzed is increasing drastically, making manual segmentation an even less efficient option in clinical practice. Additionally, human delineation might not be sufficiently reproducible. Tools for automated segmentation are therefore needed.

Manuscript received August 12, 2008; revised November 09, 2008. First published January 06, 2009; current version published June 24, 2009. *Asterisk indicates corresponding author.*

*I. Išgum is with the Department of Radiology, Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands (e-mail: ivana@isi.uu.nl).

M. Staring, M. A. Viergever, and B. van Ginneken are with the Department of Radiology, Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands (e-mail: max@isi.uu.nl; bram@isi.uu.nl).

A. Rutten and M. Prokop are with the Department of Radiology, University Medical Center Utrecht, 3484 CX Utrecht, The Netherlands (e-mail: a.rutten@umcutrecht.nl; m.prokop@umcutrecht.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2008.2011480

Computerized segmentation methods that make use of registration are gaining popularity. The registration algorithms that these methods rely on are generally applicable to a wide range of medical data (see [1]–[4]) and are automatic. Improvements in algorithms and increased computer power enable the registration of large image volumes in a reasonable amount of time. In registration, the spatial correspondence between voxels in two images, often called the fixed and the moving image, is determined. The moving image is transformed to the fixed image so that both appear to be similar. Similarity is often defined by the intensity correspondence of the two images. Examples of popular similarity metrics are the sum of squared differences (SSD), the (normalized) correlation ratio, and the mutual information (MI) measure.

There is a number of ways in which registration can be used for segmentation. A detailed overview can be found in [5]. The simplest way is by registering one manually labeled image directly to the image to be segmented (the target image). This manually labeled image, often called the atlas, can be selected in a number of ways, for example randomly or by visual inspection according to some predefined criterion, or it can be created artificially. To obtain a segmentation of the target image, the manual labeling of the atlas is transformed using the mapping determined during the registration. This process is often called label propagation and has been used in many studies, e.g., [3], [6]–[10].

Alternatively, instead of selecting only one manually labeled image as an atlas, the atlas can be constructed from a larger number of images. From a given set of images, one is chosen to be the reference image. All remaining images from the set are registered to this reference. For example, in [11] and [12] multiple images of the same patient are obtained. After all images are registered to the reference, they are subsequently averaged. Therefore, the final result shows high signal to noise ratio allowing image segmentation to be performed. The segmented image is then used as the atlas. Another example where an atlas has been created from multiple images is presented in [9] and [13]–[15]. Here, the images were scans of different patients. In both these atlas creation approaches one needs to decide how to choose the reference image. This is done either by averaging the transformed images and their manual labels, or by applying the average transformation to the reference image and its segmentation. To acquire a stable atlas, an iterative atlas generation scheme has been used (see [3], [16]–[19]). The output of one atlas generation step is used as input in the following step.

The advantage of the approaches discussed so far is that once an atlas has been generated, only a single registration from the

atlas to the target image is required to obtain a segmentation. However, it is not guaranteed that the atlas is a good representation of a complete population. Also, the described approaches depend on the success of a single registration. If the registration fails, so will the segmentation. To overcome these potential problems and to make the method more robust, Rohlfing *et al.* [3] and Heckemann *et al.* [20] proposed to register a number of manually segmented images, referred to as atlases, to the target image. This circumvents the need for a single intermediate representation of multiple presegmented images. This multi-atlas-based segmentation offers a number of options for obtaining the segmentation of the target image. For example, only the result of the most successful registration can be used for label propagation, where the criterion for most successful can be derived from the registration measure (e.g., normalized mutual information) or the deformation field (e.g., small deformations are preferable). One can also use a subset of the most successful registrations or use all registrations. Typically, label propagated segmentations are then averaged, but other combination rules such as voting are also possible. This process is called decision fusion.

In [3] these different atlas-segmentation strategies were compared. It was shown that segmentation results based on multiple registrations were more accurate than those that used only a single registration.

In this work, we also use multiple registrations of manually labeled images to arrive at a segmentation. Instead of combining all or a subset of the registrations with a single, global rule, we estimate the success of the different registrations *locally* and use this to determine spatially varying weights. Additionally, the possibility of reducing the number of atlases is investigated, because performing multiple registrations in order to arrive at a single segmentation is a computationally time consuming approach. A procedure is proposed that takes the given set of atlases and selects a subset that gives the best segmentation result. The method is tested on the segmentation of the heart and aorta in computed tomography (CT) images of the thorax. The segmentation of those anatomical structures is important for the detection and analysis of cardiac and vascular abnormalities, respectively. For example, in order to be able to automatically detect coronary calcifications as described in [21], automatic segmentation of the heart and the aorta is crucial. Precise segmentation of the aorta in CT scans is also needed for planning biopsies of lymph nodes in the vicinity of the aorta to prevent aortic puncture [22]. Because both the registration algorithm and the atlas selection procedure have many parameters, it is difficult to directly compare our results with those of other studies. Therefore, we implemented some of the previously proposed methods and tested them on the given data.

II. Methods

First, in Section II-A the used registration algorithm is described. In Sections II-B and II-C the proposed segmentation algorithm is presented, followed by a description of previously proposed atlas-based segmentation methods in Section II-D.

A. Registration

In order to be able to propagate the labels from an atlas to an unseen (target) image, a registration is needed between the two. The atlas A is always chosen to be the moving image, and is transformed to the unseen fixed image U. The registration problem is formulated as an optimization problem, in which the similarity between the fixed and moving image is maximized with respect to the transformation $\boldsymbol{u}$

$$\hat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} \mathcal{C}\left[\boldsymbol{u}; U(\boldsymbol{p}), A(\boldsymbol{p})\right] \tag{1}$$

where $\hat{\boldsymbol{u}}$ is the optimal transformation making $A(\boldsymbol{u}(\boldsymbol{p}))$ spatially aligned to $U(\boldsymbol{p})$, $\boldsymbol{p} = (x, y, z)$ denotes a voxel in the image, and $\mathcal{C}$ is an appropriate cost function.

As a cost function the negative mutual information was used, following the implementation of Thévenaz and Unser [23]. For the transformation initially an affine transformation was used to get a global alignment of the two images. Subsequently, a nonrigid registration was applied to account for local differences between the atlas and the target image. This nonrigid transformation was modelled by *B*-splines. The employed registration framework is largely based on the papers of Rueckert *et al.* [1] and Mattes *et al.* [2].

For the optimization of $\mathcal{C}$ in (1) an iterative stochastic gradient descent optimizer is used. In each iteration a step is taken towards the minimum. The direction of this step is based on the derivative of $\mathcal{C}$ to the transformation parameters. The derivative is calculated based on a small subset of the image samples, randomly chosen every iteration, in order to speed up the registration [24]. A multiresolution strategy is taken to avoid local minima. To this end a Gaussian pyramid is employed, using a subsampling factor of two. Also, a multigrid approach is used for the nonrigid registration, meaning that the registration is started with a coarse *B*-spline control point grid, which is refined in subsequent resolutions.

The implementation of the applied registration algorithm can be found online.[1] This software package is based on the Insight Segmentation and Registration Toolkit (ITK), which can be found online.[2]

For all atlas-based segmentation methods described in the following sections, the same registration framework with equal parameter settings was used.

B. Label Propagation and Weighted Decision Fusion (WDF)

During the registration, a transformation is determined which transforms the moving image to the fixed image. When both the atlas and the target images are acquired by the same modality, regions of interest have the same intensity ranges in both images. Additionally, if large pathologies are not present to affect the intensity and texture of the regions of interest, we can assume that in the ideal case the transformed moving image would be equal to the target image and the difference between them would be a zero image. In reality, registration does not perfectly align the two images. The difference between the transformed moving image and the target can be inspected to evaluate the success of the registration. The closer the value at some point in the difference image is to zero, the better the registration at that point. The difference image usually consists of a wide range of

[1]http://www.elastix.isi.uu.nl

[2]http://www.itk.org

gray values, which means that the registration was able to determine the transformation aligning some parts of the image better than others. We propose that a registration which has been more successful contributes more to the segmentation result than one which was less successful. Using a difference image, this success can be evaluated locally for each registration.

Formally, the method works as follows. We have N atlas images $A_1, A_2, \ldots, A_N$ with corresponding manually segmented binary images $S_1, S_2, \ldots, S_N$. Let U be the target image, and S its unknown binary segmentation. Each atlas was registered to the target image. This resulted in N transformations $\boldsymbol{u}_i$, describing transformations from A_i to U. These transformations were used to propagate labels S_i to the target U. Each label propagation provides an opinion about the label of a particular voxel. In order to determine the relative importance of each opinion, the absolute difference D_i between the transformed moving atlas and the target image was computed

$$D_i(\boldsymbol{p}) = \left| A_i\left(\boldsymbol{u}(\boldsymbol{p})\right) - U(\boldsymbol{p}) \right|, \quad \forall i. \tag{2}$$

Subsequently, D_i was convolved with a Gaussian kernel g_{σ_1} at scale σ_1 to obtain a smoothed local estimate of the registration success [25]. To determine how much a propagated label in each transformed image should contribute to a segmentation, weights λ_i were assigned according to

$$\lambda_i(\boldsymbol{p}) = \frac{1}{D_i(\boldsymbol{p}) * g_{\sigma_1}(\boldsymbol{p}) + \epsilon} \tag{3}$$

$\forall i$, where ϵ is a small value to avoid division by zero, in this paper set to 0.001. The weight image λ_i is inversely proportional to a value in the absolute difference image, so large values in the absolute difference image results in small weights, and vice versa.

The probabilistic label was determined by a weighted average of the transformed binary segmentations $S_i(\boldsymbol{u})$

$$S_p(\boldsymbol{p}) = \frac{1}{\sum_{i=1}^{N} \lambda_i(\boldsymbol{p})} \sum_{i=1}^{N} \lambda_i(\boldsymbol{p}) S_i\left(\boldsymbol{u}_i(\boldsymbol{p})\right). \tag{4}$$

In this way for each voxel a value between zero and one was determined that corresponds to the probability that voxel is inside the object to be segmented. To obtain a binary segmentation S of U, S_p was first blurred with a Gaussian kernel with width σ_2

$$S_{\sigma_2}(\boldsymbol{p}) = S_p(\boldsymbol{p}) * g_{\sigma_2}(\boldsymbol{p}) \tag{5}$$

and subsequently thresholded at 0.5. The generated binary image S may contain some isolated voxels or groups of voxels at the border of the segmented object. To remove those, 3-D component labelling was performed, and only the largest component was retained.

Note that this algorithm has two free parameters: the Gaussian kernel sizes σ_1 and σ_2. Also note that the presented method processes 3-D CT data. Both the registration and the label propagation were performed based on volumetric data.

C. Atlas Selection (WDFS)

The WDF method uses all atlas scans. It is however not likely that all these atlases are equally useful to perform a certain segmentation task. Moreover, registration is a time consuming operation, so reducing the number of atlases, thus reducing the number of registrations that must be performed to segment a single target image, is advantageous. We propose to select a subset of atlases with a method similar to what is known as sequential forward selection (SFS) in statistical pattern recognition [26], [27].

In SFS, features are chosen in a stepwise fashion to give the best classification result. This means that at each step, the feature giving the best classification performance in combination with the already chosen features is selected. Likewise, atlases were selected in a stepwise fashion to give the best segmentation performance with already chosen atlases. The segmentation performance was measured as an overlap between the automatically computed segmentation and the corresponding manual segmentation expressed in terms of Tanimoto coefficient J

$$J(S, MS) = \frac{||S \cap MS||}{||S \cup MS||} \tag{6}$$

where $||\cdot||$ denotes cardinality, $\cap$ intersection, and $\cup$ union, S is the automated, and MS the manual segmentation of the target image.

D. Previously Proposed Atlas-Based Segmentation Methods

Registration results may vary depending on the exact registration parameters and experiment setup. Therefore, to reliably compare the performance of WDF and WDFS to several previously published atlas-based segmentation approaches, these were implemented and the parameters were chosen equal.

1) Atlas-Based Segmentation With a Single Best Atlas (SBA): In [3] two different atlas-based segmentation approaches with a single atlas were used. In the first approach, from a set of manually segmented images, a single one was chosen as an atlas by visual inspection. In the second approach, for each target image a single atlas most similar to the target was chosen based on several predefined criteria. In our implementation, from the set of N manually segmented images, we chose the atlas which is on average able to deform the best to all target images. To select this single atlas, the difference images D_i were observed. Only the volume of interest and voxels nearby it were evaluated. To precisely define this region in each image, distance transform maps were calculated with the segmentations obtained by each registration, and a threshold of 20 voxels was set on the distance. Subsequently, the average value in the difference image was computed over all target images, and for each atlas scan. The atlas giving the smallest sum of absolute values was selected as the one giving the best performance. The selected atlas was registered to all target images and its labels were propagated to obtain segmentations in the target images.

2) Average-Shape Atlas-Based Segmentation (ASA): A single atlas A_j was randomly chosen from the set of N atlases. The remaining $N-1$ atlases were registered to A_j, which resulted in $N-1$ transformations $\boldsymbol{u}_i$. The average-shape atlas ASA was constructed by taking the average of the deformed atlases

$$\mathrm{ASA}(\boldsymbol{p}) = \frac{1}{N-1} \sum_{i=1}^{N-1} A_i\left(\boldsymbol{u}_i(\boldsymbol{p})\right).$$

Segmentation of the target image was computed by registering the single average-shape atlas ASA to the target image, and propagating its segmentation. This approach has been used in [13].

3) Multi-Atlas-Based Segmentation With Averaging as Decision Fusion (ADF): Label propagation was performed the same way as in WDF. To obtain the probabilistic segmentations, decision fusion was performed by averaging results of each transformation. To be precise, in (4) we set $\lambda_i = 1, \forall i$. Finally, to obtain a binary segmentation, S_p was blurred with Gaussian kernel width σ_2 and thresholded at 0.5. A similar implementation is described in [3].

III. Experiments and Results

Performance of the proposed method was tested on the segmentation of the heart and the aorta in CT scans of the thorax. The method was also compared to the atlas-based segmentation approaches described in Section II-D.

A. Data

In the experiments 29 low-dose, noncontrast enhanced CT scans of the thorax were used. The scans were obtained from asymptomatic subjects as part of a lung cancer screening trial. CTs were acquired on a Mx8000 IDT scanner from Philips Medical Systems (Cleveland, OH) with 16×0.75 mm slice collimation and 30 mAs at 120 kVp or 140 kVp depending on subject size. All scans were realized in about 12 s. They were performed in full inspiration after the appropriate instructions were given. No spirometric control, nor respiratory belt were used. No contrast material was administered. All scans were reconstructed to a 512×512 matrix and a moderately soft kernel (Philips B). CT data was acquired at a bit depth of 12 bits/pixel and no further downsampling of bits/pixel was performed. During the automatic segmentation the original pixel values (Hounsfield units) were used, and therefore window and level settings do not play a role in the method. The smallest field of view was used that included the outer rib margins at the widest dimension of the thorax. This resulted in an in-plane resolution between 0.6 and 0.7 mm. Slice thickness was 1 mm with 0.7 mm increment. The 29 scans were randomly chosen from the set of about 500 baseline screening images. They were further randomly divided into a set of 15 atlases and a set of 14 target images.

In addition, the performance of the proposed method was tested on data containing abnormalities. For this purpose additional experiments were performed on scans from the lung cancer screening trial which contained larger abnormalities and on six scans of patients with interstitial lung disease (ILD). The latter scans were acquired on multislice scanners (Brilliance-16P, Brilliance-40, Brilliance-64, and Mx8000 IDT 16, Philips Medical Systems, Cleveland, OH). Collimation varied between 0.625 mm on the 40- and 64-slice scanners and 0.75 mm on the 16-slice scanner. Images of 0.9 mm thickness or 1 mm thickness on, respectively, the 40-/64-slice scanner and the 16-slice scanner were reconstructed every 0.7 mm. Exposure settings were 120 kVp and between 100 mAs and 170 mAs, depending on a scanner and patient size. No contrast material was administered, no ECG synchronization was performed.

B. Manual Segmentation

The heart and the aorta were segmented by two medical students who were trained and supervised by a radiologist, and have worked independently. The 15 atlases were segmented by one student, and the target images by both of them. The results of the observer who segmented all images were used as the reference standard; the results from the other observer were used to compute an estimate of the interobserver variability for comparison with the automatic results.

Manual segmentation was performed in transverse slices using software specifically developed for this study. The vertical range for the segmentation of the aorta was determined by the top of the aortic arch at the top of the scan, and the apex of the heart at the bottom. For the heart segmentation the vertical range was defined by the bifurcation of the pulmonary artery at the top, and the apex of the heart at the bottom. To determine the top slice of the aortic arch and the apex of the heart, a sagittal view was used. Manual segmentation was performed with center level set to -50 HU and window width set to 400 HU.

The observer manually set a large number of points on the border of the aorta and the heart. Straight lines were drawn between those points. The top, bottom and typically every fifth to tenth slice in between them were manually segmented. The boundary was linearly interpolated in the remaining slices. The observer could subsequently move, add and delete points to correct the interpolated contours. Afterwards, binary images denoting the segmented volumes were computed.

Because the exact borders of both the heart and the aorta are often barely visible, an observer needs to inspect neighboring slices and surrounding anatomy to determine the precise position of the border. This is especially the case in the basal portions of the heart and in parts of the ascending aorta. In addition, the segmentation protocol instructed observers to perform the delineation in transverse slices. For both reasons there are some slight inconsistencies in sagittal and coronal views of the manual segmentations.

Typically, manual segmentation of the heart took about 90 min, and of the aorta about 60 min.

In additional experiments, which show the effectiveness of the method on data with abnormalities, manual segmentation was performed by a single observer. Contiguous slices with pathology were chosen by visual inspection. From the lung cancer screening trial on average 12 slices per scan were segmented, and an average of 11 slices per scan from the ILD data set.

C. Registration

All 29 images were down-sampled with a factor two in each direction using block averaging (the mean of eight voxels becomes the new voxel value) in order to reduce the required computer memory and computational load. The 15 atlases were registered to each of the 14 target images. The registration parameters were determined in a set of pilot experiments by visual inspection of the registration results.

For the affine registration four resolutions were used, in each of which 512 iterations of the stochastic gradient descent optimizer were performed. The derivative of the mutual information was calculated based on 2048 image samples, randomly chosen

TABLE I
AORTIC SEGMENTATION: AVERAGE J AND ITS STANDARD DEVIATION BETWEEN THE REFERENCE STANDARD AND THE AUTOMATED METHODS

	average J	standard deviation
2nd observer	0.8468	0.0244
WDFS	0.7763	0.0400
WDF	0.7688	0.0417
ADF	0.7304	0.0429
ASA	0.6744	0.0356
SBA	0.6735	0.0396

every iteration. For the nonrigid *B*-spline registration five resolutions were used. The *B*-spline grid spacing used in these resolutions was 64, 64, 32, 16, and 8 voxels, respectively. The optimizer performed 256 iterations in each resolution. To estimate the derivative of the mutual information 4096 image samples were used, again randomly chosen every iteration. For both affine and nonrigid registration 32 histogram bins were used.

With these settings, a single registration typically required approximately 15 min on a standard high-end PC.

D. Segmentation

Each binary image S_i containing the segmentation of the heart or the aorta was transformed to each target image using the transformation $\boldsymbol{u}_i$ obtained by the corresponding registration. Before determining the weights for a decision fusion, difference images D_i were convolved with a Gaussian kernel g_{σ_1}. After experimenting with different kernel sizes, the best results were obtained with $\sigma_1 = 0.5$ voxels for the aorta, and $\sigma_1 = 2$ voxels for the heart. The probabilistic segmentation S_p of each target image was computed by fusing the decision of the 15 transformations. The binary segmentation S was obtained by blurring the probabilistic segmentation S_p with a Gaussian kernel width $\sigma_2 = 1$ voxel, and subsequently thresholding at the probability 0.5. These parameters were chosen equal for all segmentation methods.

To obtain segmentation of the images in their original size, the segmentation results S were super-sampled to the original resolution.

E. Evaluation

For each target image, the Tanimoto coefficient between the reference standard and the automated segmentation was computed for all implemented methods. Subsequently, the average and standard deviation of Tanimoto coefficients were determined for all methods.

Average Tanimoto coefficients and corresponding standard deviation are listed in Table I and in Table III, for the aortic and the cardiac segmentation, respectively.

To compare the performance of the WDF(S) with the other methods, a two tailed paired t-test was performed. In the Table II and Table IV p-values are listed for the aortic and cardiac segmentation, respectively. The proposed method both without (WDF) and with atlas selection procedure (WDFS) performed significantly better than any other automatic method in both segmentation tasks. In the case of cardiac segmentation the difference between the second observer versus WDF(S) and ADF was not significant. Additionally, ADF was also significantly better than the automated atlas-segmentation approaches

TABLE II
AORTIC SEGMENTATION: SIGNIFICANCE OF DIFFERENCE BETWEEN J FOR THE VARIOUS TESTED METHODS; p-VALUES FROM A TWO TAILED PAIRED t-TESTS

	p value
WDFS vs. 2nd observer	< 0.001
WDFS vs. WDF	0.06
WDF vs. ADF	< 0.001
ADF vs. ASA	< 0.001
ADF vs. SBA	< 0.001
ASA vs. SBA	0.93

TABLE III
CARDIAC SEGMENTATION: AVERAGE J AND ITS STANDARD DEVIATION BETWEEN THE REFERENCE STANDARD AND THE AUTOMATED METHODS

	average J	standard deviation
2nd observer	0.8794	0.0220
WDFS	0.8847	0.0331
WDF	0.8770	0.0362
ADF	0.8770	0.0362
ASA	0.8066	0.0593
SBA	0.8021	0.0344

TABLE IV
CARDIAC SEGMENTATION: SIGNIFICANCE OF DIFFERENCE BETWEEN J FOR THE VARIOUS TESTED METHODS; p-VALUES FROM A TWO TAILED PAIRED t-TESTS

	p value
WDFS vs. 2nd observer	0.61
WDF vs. 2nd observer	0.83
ADF vs. 2nd observer	0.70
WDFS vs. WDF	0.008
WDF vs. ADF	0.007
ADF vs. ASA	< 0.001
ADF vs. SBA	< 0.001
ASA vs. SBA	0.75

using a single atlas (ASA and SBA). Finally, no significant difference was found between the ASA and SBA approach. These conclusions are valid for both segmentation tasks.

To analyze the distribution of the results, box-and-whisker plots were made [28]. They are shown in Fig. 1 and Fig. 2 for the aortic and the cardiac segmentation, respectively. For aortic segmentation, there is a clear difference between medians and distributions between the second observer and computerized methods. WDF and WDFS have a similar distribution which is narrower and has higher medians than those of the other automated approaches. In the case of cardiac segmentation, the second observer and the atlas-based segmentations with decision fusion (WDFS, WDF, ADF) have comparable performance, with higher medians and narrower distributions than the average-shape or the best single atlas segmentation approaches.

Figs. 3 and 4 each show two slices of the aortic segmentation result in two different subjects. Fig. 5 shows one slice of the cardiac segmentation in two different subjects. In both segmentation tasks first example shows image where WDF resulted in the highest J, and the second example image where J was lowest. In case of aortic segmentation, those images correspond to outliers in the box-and-whisker plot for WDF in Fig. 1.

To evaluate if selecting a subset of atlases could reduce computation time, but keep the performance comparable, atlas selection was performed (WDFS). Selection was tested on the complete set of target images. The results for the aortic segmentation

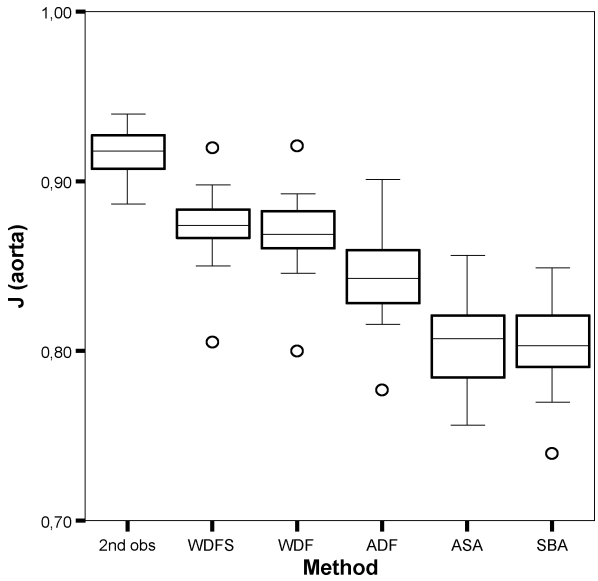

Fig. 1. Box-and-whisker plots for the aortic segmentation methods. Boxes are interquartile range. The line within the box is the median, the lines projecting out of the box contain the adjacent values which are not more than 1.5 times the height of the box. All remaining points are outliers.

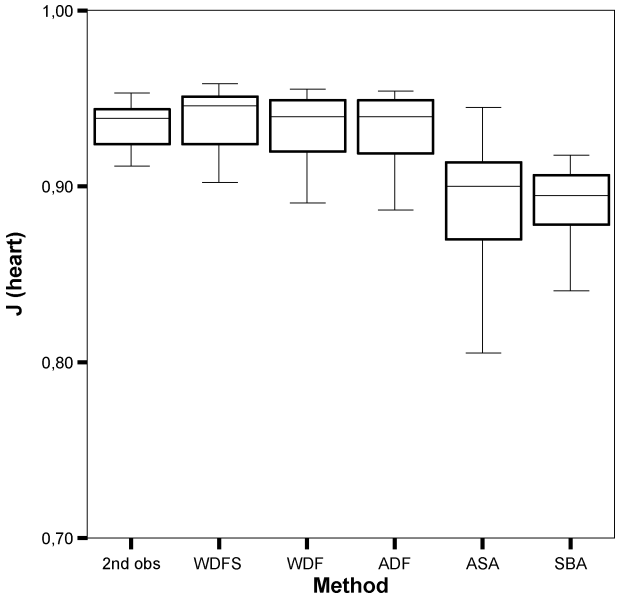

Fig. 2. Box-and-whisker plots for the cardiac segmentation methods. Boxes are interquartile range. The line within the box is the median, the lines projecting out of the box contain the adjacent values which are not more than 1.5 times the height of the box.

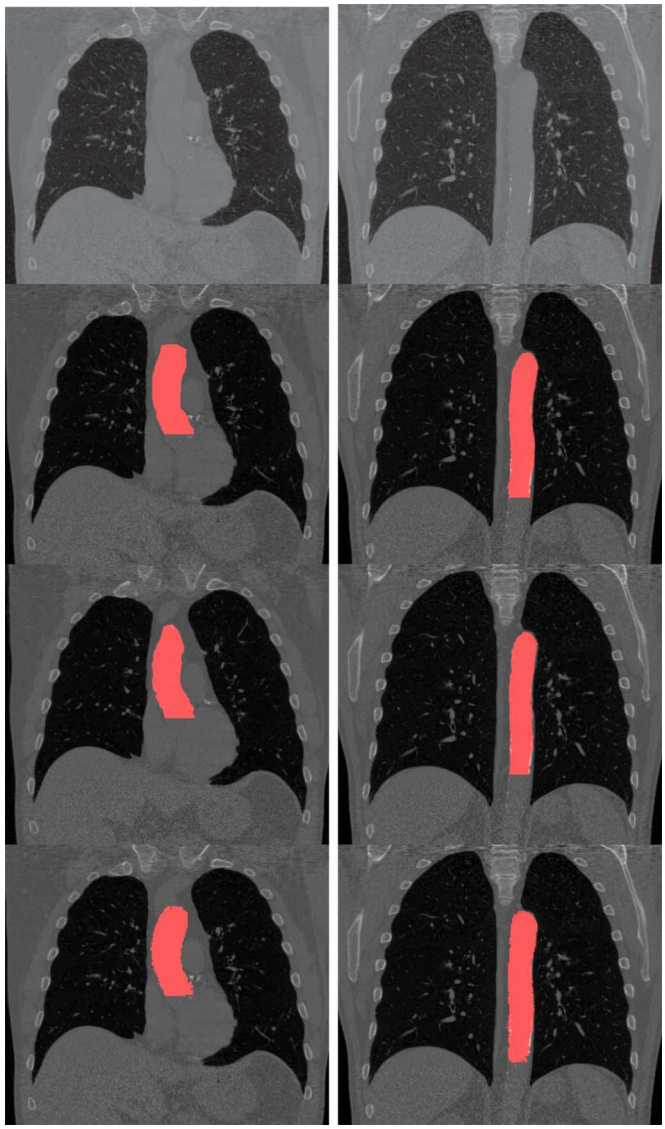

Fig. 3. Aortic segmentation in two coronal slices in the image where the automated method resulted in the highest J. The first row shows the gray value images, the second shows the reference standard, and the third row presents the segmentation of the second observer. Results of the automated segmentation by WDF are given in the last row.

are shown in Fig. 6, and for the heart in Fig. 7. The top plots in both figures show box-and-whisker plots of the average Tanimoto values over all target images for each atlas that could be selected. This means that when the first atlas was selected, performance of all 15 atlases was tested. The one with on average the highest Tanimoto coefficient was selected. In the next stage, the performance of the remaining 14 atlases each in combination with already selected atlas was computed, and the best one was selected and added, and so on. As expected, performance increased with an increasing number of atlases. The maximum was reached when eight atlases were selected, and by adding more atlases performance slightly dropped. The bottom plots in Figs. 6 and 7 show the distribution of Tanimoto coefficient for all target images in each atlas selection step.

The selected atlases were not the same in the two segmentation tasks. On average the Tanimoto coefficient obtained with eight selected atlases was slightly higher in case of the cardiac segmentation compared to the performance of the second observer, but that difference was not significant. However, the difference between WDF and WDFS was significant. This is caused by the fact that improvements of WDFS versus WDF were small, but consistent. In the case of aortic segmentation, the second observer was still significantly better than WDFS, and the difference between WDFS and WDF was not significant.

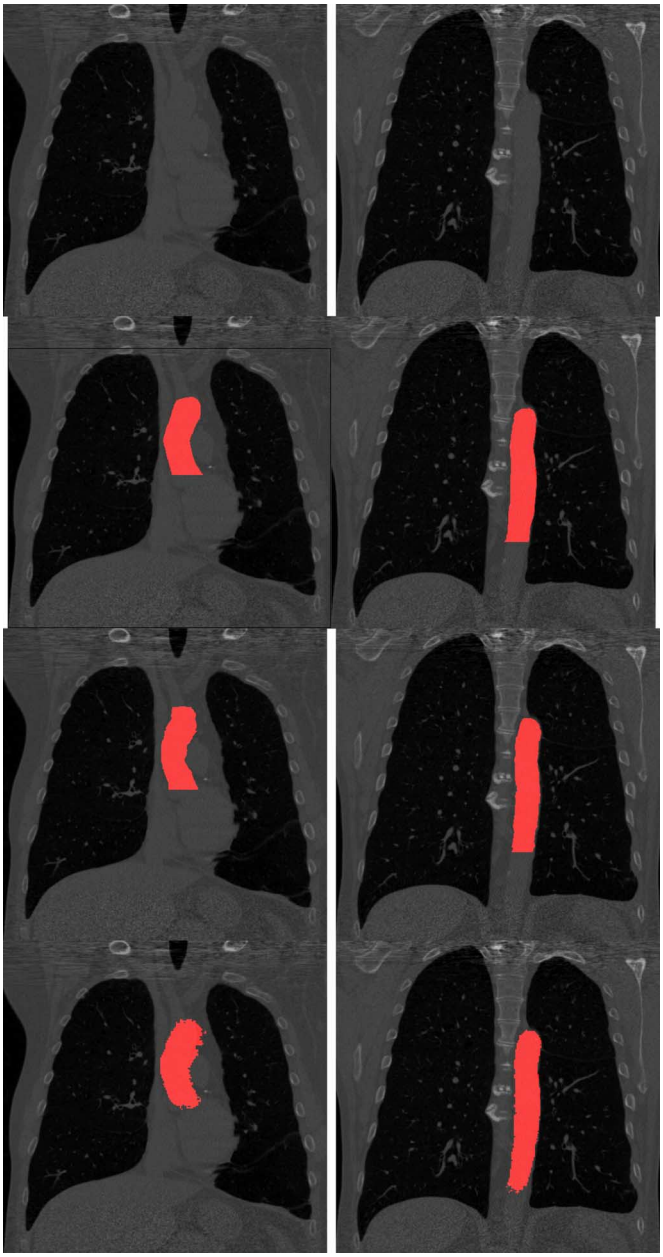

Fig. 4. Aortic segmentation in two coronal slices in the image where the automated method resulted in the lowest J. The first row shows gray value images, the second row shows the reference standard, the third row presents second observer segmentation and the last one shows results of the WDF. Note substantial variability between the observers.

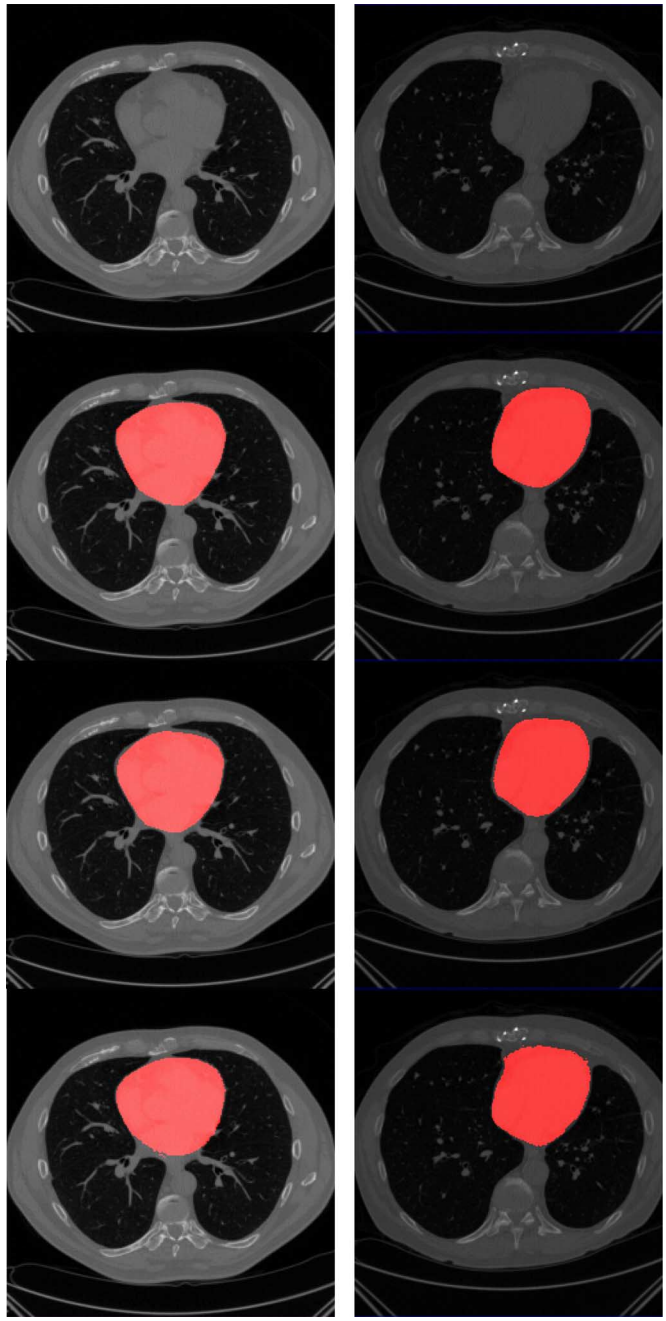

Fig. 5. Cardiac segmentation in transverse slice of two different images. In the first column is image which segmentation had the highest J, and in the second column is example of the lowest J. First row shows gray value images, second row shows the reference standard, and the third row presents second observer segmentation. Results of the WDF are shown in the last row.

To check if the proposed segmentation method would be robust to larger abnormalities in the chest scan, additional experiments were performed. First, from the screening trial four additional scans were chosen containing 1) calcified ascending aorta with substantial movement artifacts, 2) metal clips in the heart and around the sternum, 3) metal clips in the left lung from an operation, and 4) calcifications in the ascending aorta and left anterior descending coronary artery with movement artifacts. Second, the method was tested on six scans showing ILD. The results are listed in Tables V and VI. Fig. 8 illustrates the results.

IV. Discussion

The presented multi-atlas-based segmentation method yields results very close to those of an independent human observer. It is shown to be robust, in the sense that gross failures of the segmentation never occurred, although it must be noted that our tests were performed on a relatively small set of target images. In both segmentation tasks, even for the case with the lowest Tanimoto coefficient, results were good, and comparable to the reference standard when judged by visual inspection.

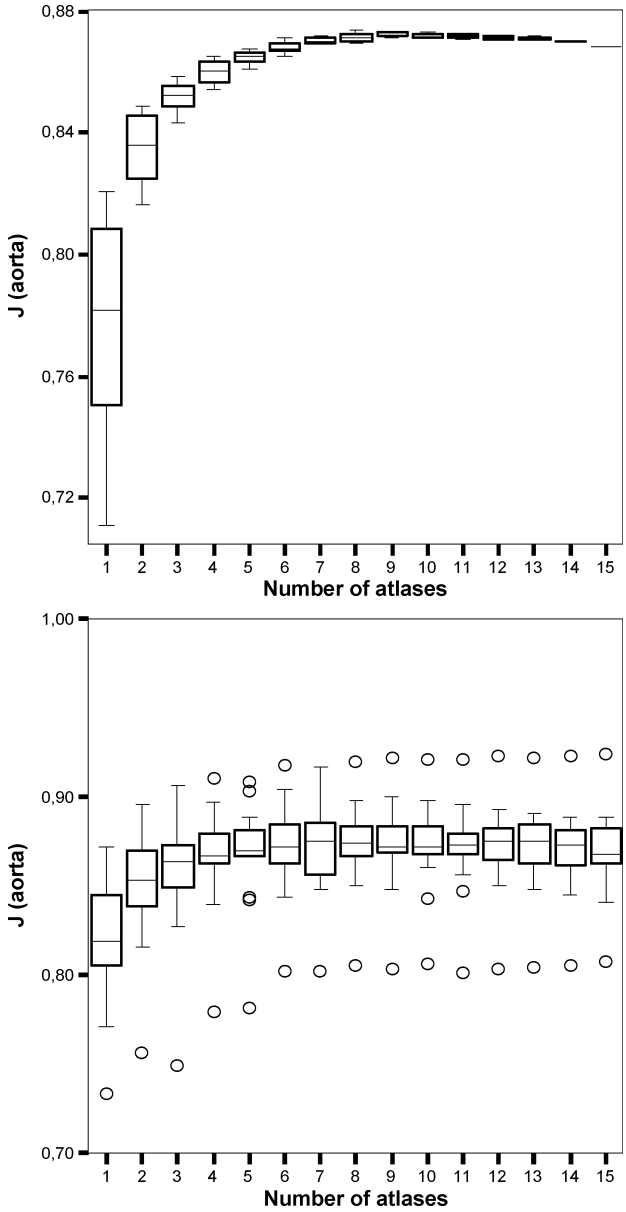

Fig. 6. Box-and-whisker plots of the atlas selection for the aortic segmentation. On the x-axis the number of selected atlases is shown. Top: The y-axis plots the distribution of the average J between WDFS and the reference standard segmentation over all target images. Bottom: On the y-axis, the distribution of the J for all target images for the selected atlases are shown. Boxes are interquartile range. The line within the box is the median, the lines projecting out of the box contain the adjacent values which are not more than 1.5 times the height of the box. All remaining points are outliers.

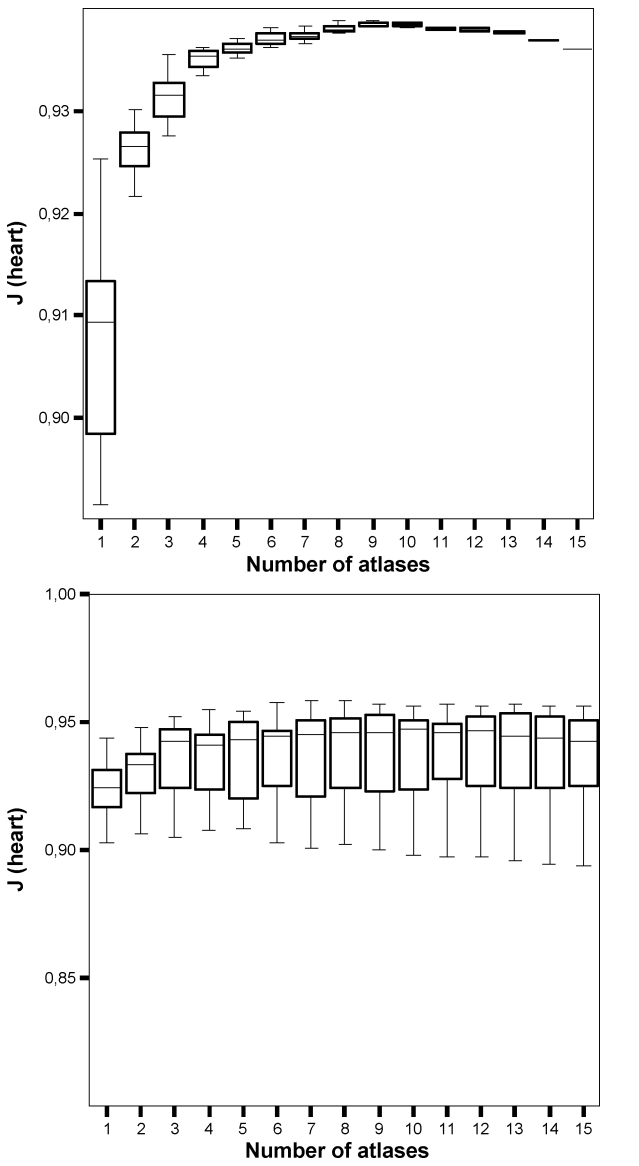

Fig. 7. Box-and-whisker plots of the atlas selection for the cardiac segmentation. On the x-axis the number of selected atlases is shown. Top: The y-axis plots the distribution of the average J between WDFS and the reference standard segmentation over all target images. Bottom: On the y-axis, the distribution of the J for all target images for the selected atlases are shown. Boxes are interquartile range. The line within the box is the median, the lines projecting out of the box contain the adjacent values which are not more than 1.5 times the height of the box.

TABLE V
AORTIC SEGMENTATION IN IMAGES WITH ABNORMALITIES: AVERAGE J AND CORRESPONDING STANDARD DEVIATIONS BETWEEN THE REFERENCE STANDARD AND WDF

	average J	standard deviation
lung cancer screening	0.6560	0.1038
ILD	0.5679	0.2024

The method has two novelties compared to already published approaches.

First, decision fusion of the propagated labels is based on the local evaluation of registration success. The local success of the registration has been evaluated using the absolute difference between the transformed atlas and the target image. Note that a different metric, namely mutual information was computed during the registration. It would have been also possible to use the same metric for both purposes. Furthermore, other ways of computing local weights are possible [see (3)].

The second novelty is an atlas selection procedure equivalent to sequential forward selection of features. It has been shown that not all atlas images were equally useful, and that the selected subset of atlases gave better performance than the complete set. Note here that in our investigation of the possibility of atlas selection, the atlases were selected on the set of target

TABLE VI
CARDIAC SEGMENTATION IN IMAGES WITH ABNORMALITIES: AVERAGE J AND CORRESPONDING STANDARD DEVIATIONS BETWEEN THE REFERENCE STANDARD AND WDF

	average J	standard deviation
lung cancer screening	0.7015	0.1738
ILD	0.8967	0.0715

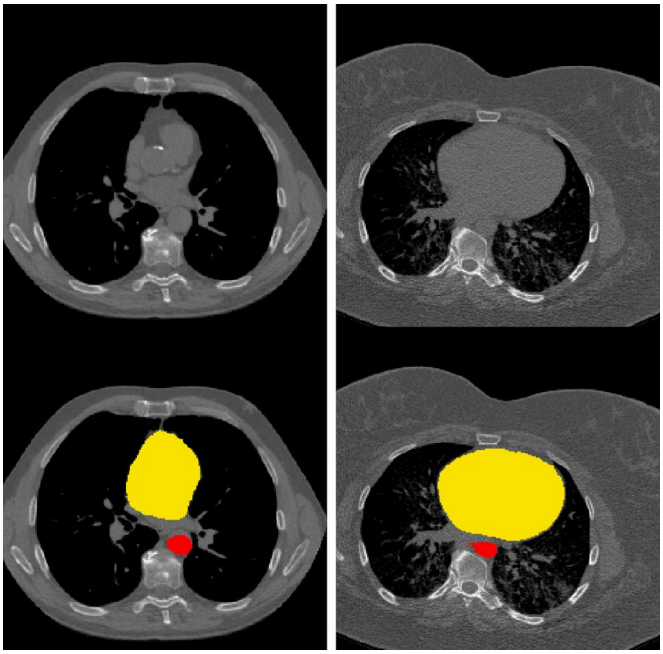

Fig. 8. Aortic and cardiac segmentation. Top, left: CT slice from the screening trial scan with calcification in the ascending aorta. Bottom, left: Segmentation results. Top, right: Slice from a CT scan showing ILD. Bottom, right: Corresponding segmentation results.

images and only that result is presented. Evaluation on an independent validation set is an important subject for future work. To perform that, a larger number of manually delineated data sets might be needed.

Simpler segmentation strategies such as flood fill algorithms or level sets would neither be able to segment the heart nor aorta in noncontrast enhanced CT scans because a strong boundary is not present.

In the presented method the best atlases were selected based on the global result. To further improve segmentation results, best atlases could be selected locally. It is possible that a better segmentation result in some parts of the target image could be achieved with a particular set of atlases. For example, some scans might be effective to register to the aortic arch, while others are most effective to find the lower part of the descending aorta. This hypothesis is supported by the fact that different atlases were selected for the heart and aorta tasks.

To additionally reduce segmentation time, the number of registrations performed, i.e. the number of atlases could be varied locally. Some parts of the target image (e.g.,, inside of the heart) can be segmented easier than others (e.g., border of the heart). Thus, for those areas where the segmentation task is easier, a smaller number of atlases may be needed. Additional registrations could be omitted locally after a few registrations have produced consistent results, and continued where the need for more opinions seems useful.

For aortic segmentation, the second observer performs better than any automated method. For cardiac segmentation on the other hand, there was no significant difference between the WDF, WDFS, and ADF versus the second observer. We believe this difference is caused by the fact that the heart has strong edges on the borders with lung tissue, which are probably effective in guiding the registration process, while for the aorta such strong edges are often not present. Therefore, the registration result was not so good in the latter case. When registration can not deform a moving image to the target image well, it is important to weigh the propagated labels locally according to the registration performance; When registration results are generally good, averaging the transformed manual segmentations is sufficient. This can also be seen when comparing the results of the WDF or WDFS and ADF methods. Although in both segmentation tasks WDF and WDFS outperformed ADF, the difference and spread of the results are greater for segmentation of the aorta than for segmentation of the heart. Box-and-whisker plots in case of the heart segmentation show comparable performance of WDF, WDFS, and ADF. However, a t-test shows that the small difference is significant. Inspecting the results for each target image showed that the difference in Tanimoto coefficient between WDF and ADF is less than 1%, but WDF always slightly outperformed ADF.

We have observed the segmentation when the best single atlas was selected from our set of atlases. There is a number of possible criteria for selecting the best atlas. Our choice was to select this image based on the registration success in the volume of interest (SBA), analogous to when registration was evaluated for WDF and WDFS. The method with the in this way chosen single atlas has given the lowest performance in terms of average Tanimoto coefficient, but comparable to the average-shape atlas method (ASA). However, box-and-whisker plots show that the J values are distributed broader in the case of ASA.

Success of the segmentation depends on registration accuracy. Therefore, the particular settings of the registration parameters (number of iterations, number of samples, number of histogram bins, step size, number of resolutions) influence the segmentation result. It is important to note that the same registration parameters were used in all experiments, because this enabled comparison between different methods. The results suggest that cases where high registration quality with a single atlas could not always be obtained profited more from the WDF approach compared to ADF.

Also, the registration results may have been influenced by the resolution of images used for registration. Even though the reference standard has been set in the full-resolution images and evaluation of the method has been performed in the original resolution, due to computer limitations, all images have been downsampled before the automatic segmentation. If more memory and faster hardware were available, the algorithm could operate on the full-resolution data and better segmentation results may be achieved.

Although the Tanimoto coefficient between the two observers is high, Figs. 3–5 show that locally substantial interobserver disagreement may occur. This is especially true at the basal por-

tion of heart and ascending part of the aorta. This is caused by the poor tissue contrast between the heart border and surrounding fat, or between the ascending aorta and surrounding tissue. An accurate registration in such areas is difficult. Furthermore, when the aorta was manually segmented, the observers were instructed not to segment the aorta below the base of the heart. However, this point is not an anatomical landmark in the aorta, and hard to find for the registration algorithm. Additionally, because manual segmentation was performed in transverse slices, sagittal views occasionally show shifts in border between neighboring transverse slices. To prevent the latter, a manual segmentation protocol could be used that instructs observers to use multiple views during manual delineation. That would however increase segmentation time substantially, and this was therefore omitted.

In the experiments the Tanimoto coefficient was used to evaluate segmentation accuracy. It is well known that this measure is dependent on the surface-to-volume ratio of the object of interest. Therefore, cardiac segmentation accuracy cannot be directly compared with the accuracy of the aortic segmentation. However, the Tanimoto coefficient is a valid measure for comparing different segmentation methods of the same structure.

Scans included in our experiments are part of a screening trial with asymptomatic participants and therefore the images typically do not contain severe abnormalities. However, frequently observed abnormalities such as aortic and coronary calcifications, aortic valve calcification, mitral valve calcium, calcifications in the tracheal wall were present in either the atlas set, the test set or in both. Moreover, two scans contained metal clips around the sternum and in the heart.

Additional experiments were performed on scans containing more severe abnormalities than the original test set from the lung cancer screening trial. As expected, performance was on average somewhat lower than in the initial set of test scans (Tables I and III versus Tables V and VI, respectively), which was mainly caused by less accurate registration result between atlas and test images. The results were still adequate though. Subsequently, the method was tested on scans showing interstitial lung disease. Results of the cardiac segmentation were slightly better than those obtained from the lung cancer screening subjects. This might be due to the fact that only slices affected by pathology were evaluated which excluded areas around the basal portions of the heart. Note here that the original set of atlases was used, that did not originate from clinical ILD data. Although the segmentation results were satisfactory given the difficulty of the image data, we do expect that better registration results and therefore higher segmentation accuracy would be achieved if atlas images obtained with the same scanning protocol as the target images were used.

To make the method robust for large anatomical abnormalities such as aneurysms, it would probably be necessary to include them in the set of atlas images. However, this would likely be possible only when the abnormalities do not severely compromise the quality of the registration. An aortic dissection is rarely visible on noncontrast enhanced CT scans, and therefore could not cause problems for the algorithm.

Scan acquisition was performed without ECG-synchronization and some subjects might not have been able to hold their breath. Thus, cardiac (Figs. 3 and 4) and possibly some breathing motion artifacts were present in the scans. Because registration applied in the algorithm is elastic, it was able to correct for size, shape and displacement of the segmentation target. However, blurring of the organ borders resulted in a lower gradient at the cardiac and the aortic border and probably did influence the registration results. Therefore, we do expect that better segmentation results would be achieved if ECG-synchronization was applied and all subjects held their breath during the entire scan acquisition.

We expect the method to be applicable to contrast enhanced CT scans. Contrast material would intensify the boundaries, especially of the aorta, and therefore the registration should be able to better align the atlas and the target image. This might lead to a better segmentation result. If the data to be segmented is a mixture of contrast and noncontrast enhanced data, it may be advantageous to include this mixture also in the training data. We have, however, obtained good results registering noncontrast to noncontrast enhanced data. Moreover, the method was applied to segmentation tasks in low-dose scans obtained in the screening program. This means that application of the same algorithm for segmentation tasks in scans obtained with higher radiation doses such as usually applied in clinical practice will likely be feasible.

We believe that the proposed method could be applied to segmentation of other anatomical structures in CT and MRI images (of the same sequence) for which other (multi-) atlas-based segmentations have already been used, such as for brain or cardiac structures.

V. Conclusion

An atlas-based segmentation method employing label propagation and spatially varying decision fusion has been presented. The method evaluates the success of the registration between the atlas and the target image locally, and on that basis a weighted decision fusion is performed. The method was tested for segmentation of the heart and aorta in CT images. The proposed method outperformed other implemented atlas-based segmentation approaches. Additionally, it has been shown that selection of a subset of atlases from the original set leads to faster and comparable segmentation results.

References

[1] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[2] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank, "PET-CT image registration in the chest using free-form deformations," *IEEE Trans. Med. Imag.*, vol. 22, no. 1, pp. 120–128, Jan. 2003.

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

[4] M. Lell, K. Anders, E. Klotz, H. Ditt, W. Bautz, and B. Tomandl, "Clinical evaluation of bone-subtraction CT angiography (SSCTA) in head and neck imaging," *Eur. Radiol.*, vol. 16, pp. 889–897, 2006.

[5] T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer, Jr., "Quo vadis, atlas-based segmentation?," in *The Handbook of Medical Image Analysis—Volume III: Registration Models*, J. Suri, D. L. Wilson, and S. Laxminarayan, Eds. New York: Kluwer Academic/Plenum, Aug. 2005, ch. 11, pp. 435–486.

[6] C. Baillard, P. Hellier, and C. Barillot, "Segmentation of brain 3D MR images using level sets and dense registration," *Med. Image Anal.*, vol. 5, pp. 185–194, 2001.

[7] A. Pfefferbaum, M. J. Rosenbloom, T. Rohlfing, E. Adalsteinsson, C. A. Kemper, S. Deresinski, and E. V. Sullivan, "Contribution of alcoholism to brain dysmorphology in HIV infection: Effects on the ventricles and corpus callosum," *NeuroImage*, vol. 33, no. 1, pp. 239–251, 2006.

[8] F. J. S. Castro, C. Pollo, R. Meuli, P. Maeder, M. B. Cuadra, O. Cuisenaire, J.-G. Villemure, and J.-P. Thiran, "Cross validation of experts versus registration methods for target localization in deep brain stimulation," in *Proc. MICCAI*, 2005, pp. 417–424.

[9] O. T. Carmichael, H. A. Aizenstein, S. W. Davis, J. T. Becker, P. M. Thompson, C. C. Meltzer, and Y. Liu, "Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 27, no. 4, pp. 979–990, 2005.

[10] P.-Y. Bondiau, G. Malandain, S. Chanalet, P.-Y. Marcy, J.-L. Habrand, F. Fauchon, P. Paquis, A. Courdi, O. Commowick, I. Rutten, and N. Ayache, "Atlas-based automatic segmentation of MR images: Validation study on the brainstem in radiotherapy context," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 61, no. 1, pp. 289–298, 2005.

[11] J. Stancanello, P. Romanelli, N. Modugno, P. Cerveri, G. Ferrigno, F. Uggeri, and G. Cantore, "Atlas-based identification of targets for functional radiosurgery," *Med. Phys.*, vol. 33, no. 6, pp. 1603–1611, 2006.

[12] M. Wu, O. Carmichael, P. Lopez-Garcia, C. S. Carter, and H. J. Aizenstein, "Quantitative comparison of AIR, SPM, and the fully deformable model for atlas-based segmentation of functional and structural MR images," *Human Brain Mapp.*, vol. 27, no. 9, pp. 747–754, 2006.

[13] I. Sluimer, M. Prokop, and B. van Ginneken, "Towards automated segmentation of the pathological lung in CT," *IEEE Trans. Med. Imag.*, vol. 24, no. 8, pp. 1025–1038, Aug. 2005.

[14] L. Zhang, E. A. Hoffman, and J. M. Reinhardt, "Atlas-driven lung lobe segmentation in volumetric X-ray CT images," *IEEE Trans. Med. Imag.*, vol. 25, no. 1, pp. 1–16, Jan. 2006.

[15] B. Li, G. E. Christensen, E. A. Hoffman, G. McLennan, and J. M. Reinhardt, "Establishing a normative atlas of the human lung: Intersubject warping and registration of volumetric CT images," *Acad. Radiol.*, vol. 10, no. 3, pp. 255–265, 2003.

[16] J. Ashburner, "Computational neuroanatomy," Ph.D. dissertation, Univ. College London, London, U.K., 2000.

[17] R. Brandt, T. Rohlfing, J. Rybak, S. Krofczik, A. Maye, M. Westerhoff, H.-C. Hege, and R. Menzel, "3-D average-shape atlas of the honeybee brain and its applications," *J. Comparative Neurol.*, vol. 492, no. 1, pp. 1–19, 2005.

[18] C. Jongen, J. P. W. Pluim, P. J. Nederkoorn, M. A. Viergever, and W. J. Niessen, "Construction and evaluation of an average CT brain image for inter-subject registration," *Comput. Biol. Med.*, vol. 34, no. 8, pp. 647–662, 2004.

[19] T. Rohlfing, R. Brandt, C. R. Maurer, Jr., and R. Menzel, "Bee brains, B-splines and computational democracy: Generating an average shape atlas," in *IEEE Workshop Math. Methods Biomed. Image Anal.*, 2001, p. 187.

[20] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[21] I. Išgum, A. Rutten, M. Prokop, and B. van Ginneken, "Detection of coronary calcifications from computed tomography scans for automated risk assessment of coronary artery disease," *Med. Phys.*, vol. 34, pp. 1450–1461, 2007.

[22] P. Taeprasartsit and W. E. Higgins, "Method for extracting the aorta from 3D CT images," in *Proc. SPIE Med. Imag.*, San Diego, CA, 2007, vol. 6512.

[23] P. Thévenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. Image Process.*, vol. 9, no. 12, pp. 2083–2099, Dec. 2000.

[24] S. Klein, M. Staring, and J. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2879–2890, Dec. 2007.

[25] L. M. J. Florack*, Image Structure*. Dordrecht, The Netherlands: Kluwer, 1997.

[26] A. Whitney, "A direct method of nonparametric measurement selection," *IEEE Trans. Computers*, vol. 20, pp. 1100–1103, 1971.

[27] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[28] J. Tukey*, Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.