Improving label fusion in multi-atlas based segmentation by locally combining atlas selection and performance estimation

Citation for published version (APA):

Langerak, T. R., Heide, van der, U. A., Kotte, A. N. T. J., Berendsen, F. F., & Pluim, J. P. W. (2015). Improving label fusion in multi-atlas based segmentation by locally combining atlas selection and performance estimation. *Computer Vision and Image Understanding*, 130, 71-79. <https://doi.org/10.1016/j.cviu.2014.09.004>

DOI:

[10.1016/j.cviu.2014.09.004](https://doi.org/10.1016/j.cviu.2014.09.004)

Document status and date:

Published: 01/01/2015

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Improving label fusion in multi-atlas based segmentation by locally combining atlas selection and performance estimation [☆]

T.R. Langerak ^{a,*}, U.A. van der Heide ^b, A.N.T.J. Kotte ^b, F.F. Berendsen ^a, J.P.W. Pluim ^a

^a Image Sciences Institute, University Medical Center Utrecht, The Netherlands

^b Department of Radiotherapy, University Medical Center Utrecht, The Netherlands

ARTICLE INFO

Article history:

Available online 2 October 2014

Keywords:

Atlas-based segmentation
Local atlas selection
Label fusion
Registration
Segmentation

ABSTRACT

In multi-atlas based segmentation, a target image is segmented by registering multiple atlas images to this target image and propagating the corresponding atlas segmentations. These propagated segmentations are then combined into a single segmentation in a process called label fusion.

Multi-atlas based segmentation is a segmentation method that allows fully automatic segmentation of image populations that exhibit a large variability in shape and image quality. Fusing the results of multiple atlases makes this technique robust and reliable. Previously, we have presented the SIMPLE method for label fusion and have shown that it outperforms existing methods. However, the downside of this method is its computation time and the fact that it requires a large atlas set. This is not always a problem, but in some cases segmentation may be time-critical or large atlas sets are not available.

This paper presents a new label fusion method which is a local version of the SIMPLE method that has two advantages: when a large atlas set is available it improves the accuracy of label fusion and when this is not the case it gives the same accuracy as the original SIMPLE method, but with considerably fewer atlases. This is made possible by better utilizing the local information contained in propagated segmentations that would otherwise be discarded. Our method (semi-)automatically divides the propagated segmentations in multiple regions. A label fusion process can then be applied to each of these regions separately and the end result can be reconstructed out of multiple partial results. We demonstrate that the number of atlases needed can be reduced to 20 atlases without compromising segmentation quality. Our method is validated in an application to segmentation of the prostate, using an atlas set of 125 manually segmented images.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

The segmentation of medical images is often required to support, amongst others, diagnosis and treatment planning. However, manual segmentation is a tedious and time-consuming task, and for this reason automatic image segmentation is currently one of the main challenges of medical image analysis. Among the many approaches to automatic segmentation, classifier combination or 'label fusion' strategies have gained popularity in the last decade. Classifier combination, a method that originates in the area of pattern analysis [11], deals with the combination of multiple observations of the same pattern. Label fusion, likewise, combines multiple estimates of the segmentation of a (medical) image. These segmentations or 'labels' are generated either manually or automatically.

Each of the estimated segmentations contains unique information that may be complementary to the information contained in other estimated segmentations and therefore a combination of multiple estimates leads to a combined result that is more accurate than each segmentation individually (see Fig. 1).

Label fusion increases the accuracy of the segmentation process, but also its robustness. Label fusion methods have been shown to better be able to handle statistical outliers than any dedicated classification strategy because they allow individual segmentations to have small errors.

In this paper, we target label fusion of segmentations that have been computed automatically in an atlas-based segmentation procedure, which derives segmentations from other, already segmented images (so-called 'atlases'). Its outline is conceptually simple (see Fig. 2): first, an image registration procedure is used to find a spatial relation between atlas and target image in the form of a transformation T that aligns the two images.

[☆] This paper has been recommended for acceptance by Albert C.S. Chung.

* Corresponding author.

Because this same spatial relation exists between the segmentations of atlas and target image, the transformation can then be applied to propagate the atlas segmentation. A more detailed introduction will be given in Section 2.

In reality, T will not perfectly propagate the atlas segmentation because it is often not possible to perfectly align two medical images. As a result there will be an error in the resulting segmentation of the target image. By registering many atlases non-systematic errors in segmentations can be corrected with the help of other segmentations [20]. In a label fusion process, all the propagated atlas segmentations are then combined to form a single segmentation of the target image. The feasibility of atlas-based segmentation using label fusion has been shown in many studies ([1–5,9,10,13,15,19,21]).

A critical aspect of atlas-based segmentation is the careful construction of the atlas set. An atlas-based segmentation strategy can only be successful if the atlas set is a good representation of the population of target images. In applications where the shape variability within the population of target images is limited, a small atlas set suffices to get good results. However, in most applications the shape variability is large and therefore a large atlas set is needed to support a robust and accurate atlas-based segmentation strategy.

The main obstacles in applying atlas-based segmentation in clinical practice are not only that this requires experts to provide manual segmentations for a large set of atlas images, but also that it requires a large amount of computation time, as every single atlas image must be registered to a target image.

Both of these requirements can often not be met in practice and because of these practical limitations small atlas sets are often used. This compromises the robustness of the label fusion process and causes the result of atlas-based segmentation to be sub-optimal in many applications.

Previously, we have proposed the SIMPLE algorithm [14], which discards atlases of which the performance is estimated to be poor before including them in the label fusion process. This method has been shown to outperform existing label fusion methods but relies on a large atlas set. Given the above-mentioned considerations, the SIMPLE algorithm may not always be applicable.

In this paper, we present a local version of the SIMPLE algorithm that has two advantages. First, when a large atlas set is available we show that the proposed approach yields an accuracy improvement compared with existing methods. Second, when only a small atlas set is available, the local SIMPLE method achieves the same accuracy as the state-of-the-art label fusion methods using a considerably smaller atlas set. We demonstrate our method in an application to prostate segmentation for the purpose of radiotherapy treatment planning. This paper is based on an earlier conference paper [16] in which we presented preliminary results.

In Section 2 we will give a short introduction to atlas-based segmentation and provide context and background for our method by referencing some of the existing literature on label fusion. In Section 3 we will define our methodology and in Section 4 we will discuss the division into regions and propose three methods for

Fig. 2. Schematic representation of atlas-based segmentation.

doing so. We will describe our data and experiments in Section 5 and give results in Section 6. Finally, conclusions and suggestions for further research can be found in Section 7.

2. Atlas-based segmentation

When a set of atlases A_i is available, atlas-based segmentation is often a successful approach to segmenting an image. An atlas $A_i = (I_i, L_i)$ can be defined as a pair of an image I_i and a label L_i that is derived from a manual segmentation. If I_{target} denotes the target image one is trying to segment, we can estimate its unknown label L_{target} by registering the atlases images I_i to the target image and then propagating each of the corresponding atlas labels L_i . Each propagated atlas label is an estimate or *classifier* of the label of the target image and all these propagated segmentations can be combined into a single estimate of L_{target} in a label fusion process [11].

An atlas-based segmentation procedure can be summarized as follows:

1. *Image registration*: a transformation function $T_i(I_i) \approx I_{\text{target}}$ that aligns the atlas images with the target image is computed by registering I_i to I_{target} .
2. *Label propagation*: given this transformation, the propagated segmentation L'_i is computed as $L'_i = T_i(L_i)$.
3. *Label fusion*: an estimate segmentation L_{est} of the ground truth segmentation L_{target} is computed by combining all labels L'_i in a label fusion procedure.

Many papers have been published on image registration, but a review of these methods falls outside the scope of this paper and we refer to the overviews that are available on this topic [17,27] and the references therein. Label propagation is a straightforward process in which atlas segmentations are propagated using the deformation field that results from the registration of the corresponding images.

2.1. Label fusion – performance estimation

This paper focuses on the label fusion process, in which propagated atlas segmentations are combined into a single segmentation of the target image. A simple way to combine segmentations is a majority voting procedure (e.g. as is done in [10]), in which each voxel in the combined image has the same value as the majority of corresponding voxels in the propagated atlas segmentations. However, in most cases a better result can be achieved by using more advanced strategies that take into account the accuracy of a propagated atlas.

A popular approach to more advanced label fusion is performance estimation: by weighing the influence of atlases based on their estimated accuracy or *performance*, one can reduce the effect of inaccurately propagated segmentations in the label fusion process and increase the effect of highly accurate segmentations. A well-known performance estimation method is the STAPLE method by Warfield et al. [26], which was extended to work for multi-class segmentations both in [26,21]. The algorithm estimates the sensitivity and specificity parameters of the performance of

Fig. 1. Simplified example of label fusion.

segmentations and that of the ground truth segmentation simultaneously using an expectation–maximization approach. Although the algorithm can be applied to label fusion of automatically generated labels, it is designed to estimate the performance of human raters and may not work optimally in the context of atlas-based segmentation [7].

When combining a large number of propagated atlas segmentations without prior knowledge on their performance, an expectation–maximization strategy will likely converge to a local optimum. In [6], a continuous version of the algorithm was presented that also works well for continuous cases, such as for example in the case of tensor images. In this method, a bias and a covariance matrix are used rather than sensitivity and specificity parameters.

Recently, an algorithm called COLLATE has been proposed, which takes into account the spatially varying performance that exists due to the fact that some regions in an image (e.g. borders with a sharp contrast) are easier to segment than other regions [3]. It estimates the consensus level for each voxel and takes this into account when locally determining rater behavior. It is shown to outperform STAPLE, especially for small numbers of atlases, but this is shown only for synthetic data.

The same authors also presented a local version of STAPLE called Spatial STAPLE, which estimates the bias of a single atlas locally by estimating a regional confusion matrix using the local cross correlation that replaces the global performance estimation during the expectation–maximization procedure [5]. This method is better able to capture the performance of atlases if this performance greatly differs over different regions of an atlas, but still suffers from the same limitations concerning a large shape variance in the atlas set.

In [2,9], the performance of atlases is estimated locally by looking at the local image similarity of the atlas image after registration and the target image. However, as noted in [2], the relation between image similarity and segmentation quality is far from optimal. In [24], Sdika proposed to use an accuracy map that is composed by pre-registering the atlas images to each other and measuring the accuracy with which this can be achieved. However, it was shown that this method only provided a significant improvement in combination with an intensity classification method, of which the accuracy is very application-specific.

2.3. Label fusion and atlas selection

An alternative to performance estimation is atlas selection [20]. In this approach, low quality segmentations are entirely discarded and the remaining atlases are combined into a single segmentation.

Aljabar et al. make a selection out of a set of 275 atlases based on the image similarity that results when these atlases as well as the target image are registered to a randomly chosen reference image [1]. This reduces the required computation time because only the target image has to be registered online. It is shown that this method outperforms selection based on meta-data in an application to brain images. However, even if a reduction in computation time can be established, the authors still assume the availability of a large atlas set, which is not a reasonable assumption in most practices.

Klein et al. apply atlas-based segmentation to the prostate [13]. Using a smaller atlas set of 50 atlases, they register all atlas images directly onto the target image. A fixed set of atlas images is then selected based on the result of the registration and these images are combined using a majority vote. Because atlas images are directly registered onto the target image, the corresponding image similarity and consequently the selection step will be more accurate than in the approach of Aljabar et al., but this strategy is not feasible for large atlas sets due to the computational burden of having to register a large atlas set.

Commowick et al. locally select atlases that are most similar to the target image, and use these local regions to create a patient-specific atlas [6]. However, to save computation time, they only register atlas images to an average atlas image which then in turn is registered to the target image. Although this is considerably more efficient, it may not be a feasible strategy when there is a large shape or appearance variability in the population of target images. Van Rikxoort et al. propose a similar local selection strategy, in which the atlas images are divided into eight blocks of roughly equal volume [19]. Then, it is locally decided which atlases must be registered based on an estimate of how much additional information the propagated atlas segmentation would locally add. The method reduces computation time, but still requires segmentations to be made available for all atlas images and is only validated on a relatively small atlas set (consisting of 39 atlases).

Cardoso et al. select atlases locally based on the local normalized cross correlation and then subject the selected atlases to the STAPLE algorithm. They present significant improvements over global selection and, more importantly, conclude that local atlas selection is more robust than global atlas selection. However, as noted before, imaging artifacts and anatomical irregularities may cause image similarity to be a sub-optimal indicator of registration and propagation accuracy.

2.4. The SIMPLE algorithm

In [14], we proposed the SIMPLE algorithm, which combines performance estimation and atlas selection in an iterative strategy that alternately estimates the performance and makes a selection of the atlases based on this estimated performance. Because discarding atlases with a poor performance will likely improve the results, this is repeated until convergence. In [15], a version of the SIMPLE algorithm was presented that estimates atlas performance locally, but still requires all atlases to be manually segmented and registered to the target image. In this paper, we present a local version of the SIMPLE algorithm that, similar to the algorithm presented in [19], semi-automatically divides a segmentation into regions that are dealt with independently. We show that, when using this method with a large atlas set, the accuracy of label fusion can be improved. When a large atlas set is not available, the method can greatly reduce the number of atlases that is required to approximately 20 atlases without compromising the accuracy of label fusion.

3. Method

In this section we give an outline of our method, which works on regions within an image, without detailing what regions are used. The division of an image into regions is discussed in Section 4.

3.1. Initialization

As an initialization of the method, we compute an initial estimate L_{est}^0 of the segmentation of I_{target} by combining all propagated segmentations L_i using a simple majority vote. In multiple studies, this very basic approach has been shown to lead to good results and we therefore assume that $L_{est}^0 \approx L_{target}$. L_{est}^0 is then divided into m regions in a procedure that we will explain in detail in Section 4.

The regions are indicated as $[L_{est}^0]^j$, with $0 < j \leq m$. It is assumed that $\forall i, j \left([L_{est}^0]^i \cap [L_{est}^0]^j = \emptyset \right)$, i.e. none of the regions overlap, and $\sum_{j=1}^m [L_{est}^0]^j = L_{est}^0$, i.e. there is no part of L_{est}^0 that is not covered in one of its regions.

All regions are then projected onto each individual segmentation L'_i . Because all atlas images have been registered to the target image, the regions $[L_{est}^0]^j$ and $[L'_i]^j$ cover the same geometrical space and the voxel sets of $[L_{est}^0]^j$ and $[L'_i]^j$ are the same.

The performance ϕ of a region $[L'_i]^j$ can then be estimated as $\phi([L'_i]^j) = DSC([L'_i]^j, [L_{est}^0]^j)$, where $DSC()$ is a function that returns the Dice Similarity Coefficient [8]. The motivation for this is that if $[L_{est}^0]^j \approx [L_{target}^j]$, it is likely that $DSC([L'_i]^j, [L_{est}^0]^j)$ approximates the unknown true performance $\phi([L'_i]^j)$, which is defined as $\phi([L'_i]^j) = DSC([L'_i]^j, [L_{target}^j])$. In the remainder of the paper we will write ϕ_i^j and ϕ_i^j short for $\phi([L'_i]^j)$ and $\phi([L_i]^j)$ while the performances of the combined segmentations are denoted as $\Phi^k = DSC(L_{est}^k, L_{target})$.

3.2. Iterative selection

After the described initialization, which can be done fully automatically, our method enters into an iterative process. In each iteration, a selection step is performed, in which those regions of the propagated atlas segmentations are discarded for which $\phi_i^j < \theta^j$, where the threshold $\theta^j = \bar{\phi}^j - \alpha\sigma^j$ depends on the mean and standard deviation of the performances of all selected atlases for region j . In [14], we give an analysis of different settings of α , but in general α can be set to $1 + \left(\frac{\tilde{S}^k}{S^0}\right)$, where $\tilde{S}^i$ is the set of selected atlases in iteration i . This means that in the first iteration all atlases with a performance that is more than two standard deviations below average are discarded, and that the selection gradually becomes more strict, as the accuracy of L_{est}^k increases. The rationale behind discarding regions of an atlas segmentation with a low performance is that these regions likely do not contribute to the end result of the label fusion process.

Once label regions have been discarded, the remaining label regions can be recombined into a single segmentation using a majority vote. Because all regions are disjoint and because they add up to a complete segmentation, a new majority vote can be computed of all partial labels on a per-voxel basis.

Based on this new majority vote, performances for each region can be re-estimated. Because the worst performing atlases have been discarded, we assume that $\Phi^{k+1} > \Phi^k$. Because the accuracy of our estimate of L_{target} has improved, in the next iteration we are better able to identify outliers that would have been missed in the first iteration. These outliers are again discarded and the entire procedure is repeated until no more outliers can be found, i.e. when the estimated performances of the remaining label regions are sufficiently close together.

3.3. Stopping criteria

During our local label fusion method, a set of selected atlas regions is maintained for each region separately. It is possible and even likely that in one region more atlases are discarded than in another region, and therefore convergence has to be detected for each region individually. If the set of labels for a specific region converges, then the algorithm will stop for that specific region but continue discarding outlying labels in the other regions. Once all regions have converged, the algorithm stops.

In addition to stopping when the method converges, we also want to stop when the method starts discarding the ‘wrong’

atlases. This occurs when the average estimated performance decreases or when the standard deviation of the estimated performance increases. In both these cases it can no longer be assumed that the estimated performance in iteration $k+1$ is better than in iteration k and therefore the method stops for a specific region in these cases.

The method is demonstrated in a two-dimensional example in Fig. 3.

4. Division into regions

By making use of local information to select local regions of an atlas, the results of label fusion can be improved [3,19]. However, a method can also be ‘too’ local: in this case contextual information is lost and the combined atlas will lose cohesion when regions are put back together. If atlas selection is done on a per-voxel basis, the result is equal to a simple majority vote. Because a per-voxel approach suffers from a lack of contextual information, a local approach in this case provides no additional benefit. Another concern of a local approach is that local results will eventually have to be recombined into a global result. This may lead to discontinuities in the global result which can be partly remedied by choosing overlapping regions. This approach is taken in both [3,19], but requires additional computation time. Care must be taken in choosing the local regions. In this section we will motivate three different ways of dividing an atlas into regions.

4.1. Blocks

The first division of segmentations is into eight blocks, which can be easily created by dividing a segmentation with three orthogonal planes through its center of gravity. The advantage of this division in regions is that it can be very easily computed. The disadvantage is that, because this division is very coarse, there is little meaningful contextual information within each region.

4.2. User-defined regions

Within a segmentation, usually different structures can be discerned by an expert. Fully automatic methods do not take expert knowledge into account, but this has the large advantage that no human intervention is needed. To still make such a method able to fuse labels locally, a techniques that is equivalent to Parzen Windows can be used [23]. However, it is impossible to realistically model human anatomy based on heuristics alone. Therefore, one of the methods we test here is a semi-automatic, user-driven method of defining regions. In [15], we presented such a method that requires an expert to define regions only once per application as a 2D radial map, which can then be converted to a spherical map that can be projected onto a segmentation. Using this method, the effort required of the expert is minimal. This method works for our application because the prostate is roughly convex and spherical, but has not been validated for other applications and it must be emphasized that it is not the purpose of this paper to propose a generically applicable method of dividing a segmentation into user-defined regions. The relevant part of the literature is the actual division into regions as is detailed in Section 4, and the main purpose of the method given in [15] is merely to make the definition of regions feasible in clinical practice.

4.3. Slices

For the data used in this paper, when creating a segmentation, experts go through an image on a slice-by-slice basis. Some slices are manually segmented, in others a contour is created based on

Fig. 3. Two-dimensional example of the regional atlas selection process for an image with four regions, with the number of the iteration step on the far left and the ground truth segmentation on the far right.

interpolation of the manual segmentations in neighboring slices. Especially in regions where the tangent of the segmentation rapidly changes, many slices will be manually segmented because interpolation is not realistic in these regions. In this case, it is not illogical that there is a bias in the atlas segmentations on a per-slice basis. Therefore we also tested a method that interprets slices as a region.

5. Data and experiments

We tested our method on a set of 125 MR-images of the prostate. Consecutive patients that were treated between mid-2007 and mid-2009 were included after giving informed consent. No criteria were used for inclusion in or exclusion from this study. All images were acquired with a 3 Tesla scanner (Philips 3T scanner, Gyroscan NT Intera, Philips Medical Systems, Best, The Netherlands) using a flex-M coil and a Balanced Turbo Field Echo (BTFE) sequence. All images consisted of $512 \times 512 \times 90$ voxels with a resolution of $0.49 \times 0.49 \times 1.0$ mm. In all images, the entire prostate was segmented by an expert. The segmentation was approved by a second expert. Fig. 4 shows slices of a few examples, where Fig. 4a shows a 'typical' prostate and Fig. 4b and c show a prostate with respectively a large and a small volume. The bottom row of Fig. 4 shows slices through the prostate in a region called the seminal vesicles, where Fig. 4d shows a 'typical' shape, whereas Fig. 4e and f show variations of the shape of the seminal vesicles. Note that these figures also show large variations in the shape and appearance of the rectum (directly below the prostate) and bladder (directly above the prostate). These figures demonstrate the difficulties that are typically encountered in our application. This is further quantified in Fig. 5, where the volumes of the region to be segmented are given. These two figures demonstrate why we need a large atlas set to be able to robustly use atlas-based segmentation. One of the experts made a division of the prostate into five different regions, based on their relevance in the practice of radiation treatment [25]. These regions are shown in Fig. 6. In this division, regions III and IV represent the areas that border on respectively the bladder and the rectum. These two regions were indicated by the expert because they border on vulnerable tissue and therefore require great accuracy. Region I represents the seminal vesicles, which are characteristic in the sense that they exhibit a much larger shape variability than the rest of the segmentation. Region V represents the apex and region II is made up of the 'rest' and has no specific meaning.

In a leave-one-out procedure, each of the 125 atlas images was segmented using atlas-based segmentation with the remaining 124 images as an atlas and the result was compared to the segmentation that was already available for that image. In each atlas-based segmentation procedure, the atlas images were registered to the target image by an affine registration followed by a non-rigid, NMI-driven B-spline registration [22] using the open source registration package Elastix ([12], available at elastix.isi.uu.nl). In Fig. 7, a histogram of all 125×124 registration results is provided, measured as the overlap between the propagated atlas segmentation and the ground truth segmentation of the target image. The transformation that resulted from each registration procedure was applied to propagate the corresponding atlas segmentation.

First, we determined the best way to divide a segmentation into regions by comparing the three methods proposed in Section 4. We performed all 125 leave-one-out experiments once for each method of division in regions. Second, to verify that the local SIMPLE method is a valid method in its own right with a result that is at least as good as existing global label fusion approaches, we compared the local SIMPLE method to other label fusion methods, including the global SIMPLE method. Finally, we investigated if we can bring down the number of atlases that is needed for multi-atlas based segmentation by using a local approach. To this end, we computed label fusion results for sets of respectively 10, 20, 50 and 124 atlases. These atlas sets were randomly chosen in an incremental fashion, such that $\tilde{L}_{10} \subset \tilde{L}_{20} \subset \tilde{L}_{50} \subset \tilde{L}_{124}$, where $\tilde{L}$ is the set of labels.

6. Results

In Table 1, results are given for global label fusion and local label fusion using respectively blocks, regions and slices. Volume overlap was measured using the DSC and the surface distance was calculated both as the Mean Surface Distance (MSD) using the method given in [18] and the Hausdorff Distance. Results are given for the entire segmentation, but also for each individual region as was given by the expert, and are averaged over all 124 leave-one-out experiments.

From this table it can be concluded that the proposed local label fusion method only performs better than a global approach with user-defined regions. This is statistically significant ($p < 0.01$ in a paired-sample t-test). The proposed algorithm did not manage to outperform a global approach when segmentations were subdivided into slices or blocks. However, for the apex and to a lesser

Fig. 4. Slices through the base of the prostate for a prostate of average (a), large (b) and small (c) size. Figures (d)–(f) show slices through part of the prostate called ‘seminal vesicles’ for a typical (d), a very elongated (e) and a very compact (f) shape of the seminal vesicles. In all figures, the delineation of the prostate is shown in white.

Fig. 5. Histogram of the volume of the ground truth prostate segmentation in our database.

Fig. 6. Division of the prostate into five regions.

extent for the vesicles, which are respectively the bottom and top regions of the segmentation in the normal direction of the slices, an approach based on a division in slices is significantly more accurate. An explanation for this is that because the manual segmentations of the atlas images are created on a slice-by-slice basis the first and last slice of the segmentations may also form a large part of the contour of the segmentation. Therefore at the outermost slices of the segmentations it is easier to identify labels with a poor quality: these labels will not exist at certain slices and yield a 0 DSC score, which is a clear outlier. It may be better to apply a slice-based division method at the outer ends of the segmentation, and to only divide the rest of the segmentation into regions, but this is left for further research. In the remainder of the paper we will assume that labels are divided into user-defined regions using the method presented in [15].

Fig. 7. Histogram of the registration results of all registrations.

Table 1

Mean and standard deviations of the accuracy of global label fusion and local label fusion using different methods of division.

	Global	Blocks	Regions	Slices
<i>DSC</i>				
Total	0.87 ± 0.068	0.86 ± 0.083	0.89 ± 0.038	0.87 ± 0.070
Region I	0.79 ± 0.092	0.81 ± 0.11	0.83 ± 0.052	0.85 ± 0.045
Region II	0.89 ± 0.063	0.88 ± 0.065	0.89 ± 0.039	0.86 ± 0.080
Region III	0.89 ± 0.067	0.90 ± 0.093	0.90 ± 0.031	0.86 ± 0.082
Region IV	0.84 ± 0.062	0.83 ± 0.062	0.85 ± 0.034	0.81 ± 0.077
Region V	0.89 ± 0.073	0.88 ± 0.081	0.90 ± 0.041	0.92 ± 0.048
<i>MSD (mm)</i>				
Total	2.26 ± 0.40	2.45 ± 0.45	2.16 ± 0.38	2.39 ± 0.47
Region I	2.43 ± 0.43	2.89 ± 0.52	2.19 ± 0.46	2.13 ± 0.40
Region II	2.14 ± 0.38	2.31 ± 0.44	2.08 ± 0.33	2.52 ± 0.52
Region III	2.19 ± 0.41	2.29 ± 0.39	2.19 ± 0.39	2.47 ± 0.45
Region IV	2.31 ± 0.37	2.44 ± 0.48	2.17 ± 0.35	2.54 ± 0.50
Region V	2.20 ± 0.46	2.24 ± 0.38	2.12 ± 0.36	2.11 ± 0.39
<i>Hausdorff (mm)</i>				
Total	9.51 ± 2.93	9.11 ± 2.91	8.98 ± 2.64	9.41 ± 3.01
Region I	10.73 ± 4.21	10.84 ± 4.37	10.12 ± 3.25	9.27 ± 3.47
Region II	9.01 ± 3.44	9.09 ± 3.29	8.83 ± 3.51	9.14 ± 3.69
Region III	9.28 ± 2.31	8.63 ± 2.78	8.77 ± 2.39	9.06 ± 2.91
Region IV	9.13 ± 2.58	8.54 ± 2.81	8.40 ± 2.13	8.79 ± 2.62
Region V	8.47 ± 2.57	7.74 ± 2.62	7.94 ± 2.24	6.92 ± 2.31

In a second experiment we compared local label fusion to other label fusion methods. We compared the proposed method to the original SIMPLE method and to a simple majority vote (MV). In addition we compared the results of our method to an atlas selection strategy (AS) similar to that used in [13,1], where the atlases are selected that best register to the target image, measured by the image similarity used in the registration. We also employed a strategy (PE) that is similar to the ones proposed in [2,9], in which atlases are weighted by their performance, which is estimated by looking at the image similarity after registration. Finally the method was compared to a reimplementation of the ALMAS algorithm [19], the STAPLE algorithm [26] and the Spatial STAPLE algorithm [3], for both of which code from the authors of those

Table 2

Mean and standard deviation of the results of 125 experiments, measured as DSC and MSD.

	DSC	MSD (mm)	Hausdorff (mm)
<i>Global methods</i>			
Majority vote	0.83 ± 0.076	2.60 ± 0.47	13.01 ± 4.39
Atlas selection	0.84 ± 0.067	2.39 ± 0.46*	11.31 ± 3.61*
Performance estimation	0.84 ± 0.062	2.42 ± 0.34*	12.89 ± 4.07
STAPLE	0.85 ± 0.067*	2.45 ± 0.68	12.46 ± 3.87*
SIMPLE	0.87 ± 0.068*	2.26 ± 0.40*	9.51 ± 2.93*
<i>Local methods</i>			
ALMAS	0.84 ± 0.070	2.38 ± 0.55*	11.38 ± 3.54*
Spatial STAPLE	0.87 ± 0.066*	2.20 ± 0.35*	10.21 ± 3.26*
Local SIMPLE	0.89 ± 0.038*	2.16 ± 0.38*	8.98 ± 2.64*

methods was available online. Parameters for these methods were tuned with the help of those authors. In the case of Spatial STAPLE, which had difficulty dealing with our full dataset, we selected the 50 atlases with the largest image similarity after registration to serve as an input.

The results are given in Table 2, measured with the DSC and the mean surface distance. Methods that significantly outperform the majority vote ($p < 0.01$ in a paired-sample t-test) are flagged with an asterisk. The values shown are the means and standard deviations.

Local SIMPLE significantly outperformed the other methods ($p < 0.01$ in a paired-sample T-test) when measured by the DSC and the Hausdorff distance and all other methods except Spatial STAPLE when measured with the MSD. Especially the difference in Hausdorff distance is remarkable but we believe this to be a result of the fact that the SIMPLE algorithm uses an atlas selection strategy that eliminates extreme outliers. That the result of a global atlas selection strategy also yields a significantly lower Hausdorff distance seems to confirm this. Based on a visual inspection of a number of experiments the large Hausdorff distance for the other methods is dominated by small areas, mostly in region I, and therefore we must emphasize that these should not be taken as a sole measure for segmentation quality. The Hausdorff distance is only included here for completeness.

Although only data from one application was evaluated, given the results of Spatial STAPLE and Local SIMPLE we conclude that a local approach can improve the accuracy of label fusion. There are two explanations for the fact that the third local method (ALMAS) does not perform significantly better than the global methods: first, this method was originally not proposed to improve accuracy but to improve computation time. Second, to be able to compare it to the Local SIMPLE method we used the same subdivision of an entire segmentation into regions for both methods. A subdivision into regions as was proposed in [19] may have given better results.

The poor performance of the atlas selection and performance estimation strategies are due to the fact that the correlation between image similarity of the registered images and the similarity of the propagated labels is weak (as was shown in [14]) due to a poor image contrast and the presence of imaging artifacts. In other applications these two strategies may perform better. When measured by the MSD these two methods seem to perform better, but this is mostly due to the fact that the performance of the majority vote seems to be much worse due to the fact that it occasionally produced 'islands' that are disconnected from the main segmentation. These islands have little influence on the DSC, but they do have an influence on the MSD.

Although Local SIMPLE outperforms all other label fusion methods, this is accompanied by a large increase in computation time (5 h for local SIMPLE compared to 16 min for global SIMPLE). In addition, the accuracy improvement has only been shown for a large number of atlases. Although large atlas sets are available in an increasing number of applications, there are still many cases

Table 3
Evaluation of different numbers of atlases used by the Local SIMPLE method.

	DSC	MSD (mm)	Hausdorff (mm)
<i>Global SIMPLE</i>			
10 atlases	0.82 ± 0.105	2.65 ± 0.60	12.37 ± 4.53
20 atlases	0.83 ± 0.114	2.57 ± 0.57	11.68 ± 3.91
50 atlases	0.85 ± 0.097	2.43 ± 0.46	10.34 ± 3.39
124 atlases	0.87 ± 0.068	2.26 ± 0.40	9.51 ± 2.93
<i>Local SIMPLE</i>			
10 atlases	0.83 ± 0.061	2.46 ± 0.35	12.00 ± 4.39
20 atlases	0.88 ± 0.071	2.23 ± 0.41	11.01 ± 3.95
50 atlases	0.88 ± 0.063	2.21 ± 0.46	9.78 ± 3.18
124 atlases	0.89 ± 0.038	2.16 ± 0.38	8.98 ± 2.64

in which only a small number of atlases can be used. The reason for using 124 atlases is twofold: first, a comparison between the different methods is more reliable when using large numbers of atlases. Second, for the global SIMPLE algorithm we noticed that a large number of atlases does improve the results (as shown in Table 3), both in terms of the average accuracy and the standard deviation of those results. This seems at odds with existing literature because other authors ([1,9,10,19,23,24], amongst others) have shown that the added benefit of additional atlases is negligible after at most 20 atlases. However, the SIMPLE method (both local and global) discards some of the atlases it receives as an input and not all of these are actually being used in the end result of the method. This explains why the number of atlases that this method needs is larger than the optimal number of atlases mentioned in existing literature. In other words: although the original SIMPLE method improves on other global label fusion methods, this improvement depends on the availability of a large number of atlases. Using a local version of the SIMPLE algorithm, not only can we improve the accuracy of label fusion, but we can also decrease the number of atlases that is needed. Because the Local SIMPLE method partially uses atlases instead of discarding them entirely we expect it to need fewer atlases than the global SIMPLE method. To investigate whether this is indeed the case in our application, we repeated our experiments for randomly chosen sets of 50, 20 and 10 atlases both with the global SIMPLE algorithm and with the local SIMPLE algorithm. Random atlas sets were chosen 10 times and the results were averaged over these random sets. Means and standard deviations are provided in Table 3.

From this table, it can be concluded that a local label fusion strategy with as few as 20 atlases does not perform significantly worse than local label fusion with 124 atlases. The results confirm that the global SIMPLE algorithm performs better when more

Fig. 8. Quartiles and median of the percentage of the input atlases that is fully or partially used.

atlases are used but that the local SIMPLE algorithm presented in this paper does not perform significantly better when using more than 20 atlases. The reason for this is that, when provided with fewer atlases, the local SIMPLE method uses its input more efficiently. Fig. 8 shows a boxplot that displays what percentage of the input atlases are fully or partially used for different sizes of the atlas set. Again, the Hausdorff distance is lower the more atlases are used and this, again, indicates that being able to deselect extreme outliers is the main reason for a small Hausdorff distance.

Local label fusion with 20 atlases is considerably faster than with 124 atlases, taking 27 min compared to over 5 h with 124 atlases on a standard desktop computer with a 2.66 GHz processor and 3.25 MB of RAM. In addition considerably less time was needed offline to register the atlas images (approximately 5 h compared to 27 h on a desktop computer).

7. Conclusions and future research

In this paper we proposed a new local label fusion method for multi-atlas based segmentation that is an extension of the previously published SIMPLE algorithm. A local approach to atlas-based segmentation in itself is not a new idea, but results show that, when using a large atlas set of 124 atlases, the proposed method can be used to achieve a significant accuracy improvement compared to state-of-the-art methods, including the original SIMPLE algorithm. However, the main purpose of this paper was to show that the benefits of the original SIMPLE algorithm such as simultaneous atlas selection and performance estimation as well as the ability to take advantage of large atlas sets [14] translate to a local approach. Our method takes a considerable amount of computation time, which may make it a less attractive option in time-intensive applications.

Therefore, we have also shown that the method does not perform significantly worse when using only 20 atlases. However, although we still find an accuracy improvement over existing methods this is no longer statistically significant and the only claim that we can therefore make is that the local SIMPLE algorithm can also be applied to small atlas sets, thereby overcoming one of the disadvantages of the original SIMPLE algorithm.

In testing our method against other global and local methods, we assumed that large atlas sets are available. We have shown that the proposed method yields an improvement over existing methods, but we have not shown that our method also outperforms other methods for small atlas sets. In clinical practice, it is often difficult to provide large atlas sets and in these cases a local approach may not outperform non-local approaches.

It must be noted that although the proposed method offers a significant improvement in our application, this may not be the case in other applications. As demonstrated in Figs. 3 and 4, our application exhibits a large variability in shape and volume. This large variability forces us to use a large atlas set in order to be able to robustly segment images in the target population: because the variability in the population is large, the variability in the atlas set must be equally large. Because the prostate is a common cancer site, many atlases are or can be made available.

In other applications, such as for example the brain, volume differences between subjects may be much smaller. As a result, a small atlas set suffices and the proposed method will likely yield a smaller improvement than is presented in this paper. In yet other applications, such as the cervix, the large shape variability of one of the target organs (the uterus) demands a large atlas set but in normal practice not enough atlases are available because cervical cancer is less common than prostate cancer. Other areas in which we expect our method to work well are the lungs and the liver: these

organs display significant variability yet are common enough to generate a sufficiently large atlas set.

In addition to our two main claims, we have confirmed the finding in other recently published papers on label fusion that a local approach, compared to a global approach, is more accurate. We have also shown that a division into regions that is based on expert knowledge facilitates a more accurate label fusion than other methods of subdivision.

In the future we plan to validate our method in other applications as well. We expect that a local approach with user-defined regions will be less effective for very small segmentations, because here it is difficult to define user-defined regions. In these cases, a slice-based division in regions may yield a better result.

In addition we plan to investigate other, automatic, methods of using expert knowledge to divide a segmentation into regions, for example based on geometric features of the segmentation. In addition, we hope to further reduce the computation time of atlas-based segmentation by locally deciding what regions to include prior to registration, as is done in [19].

Acknowledgments

We would like to thank Simon Warfield and Andrew Asman for the help in fine-tuning parameters for respectively the STAPLE and Spatial STAPLE methods and for useful discussion on the results of our experiments.

References

- [1] P. Aljabar, R.A. Heckemann, A. Hammers, J.V. Hajnal, D. Rueckert, Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy, *NeuroImage* 46 (3) (2009) 726–738.
- [2] X. Artaechevarria, A. Munoz-Barrutia, C. Ortiz-de-Solorzano, Combination strategies in multi-atlas image segmentation: application to brain MR data, *IEEE Trans. Med. Imaging* 28 (8) (2009) 1266–1277.
- [3] A.J. Asman, B.A. Landman, Formulating spatially varying performance in the statistical fusion framework, *IEEE Trans. Med. Imaging* 31 (6) (2012) 1326–1336.
- [4] A.J. Asman, B.A. Landman, Robust statistical label fusion through consensus level labeler accuracy, and truth estimation, *IEEE Trans. Med. Imaging* 30 (10) (2011) 1779–1794.
- [5] M.J. Cardoso, K. Leung, M. Modat, J. Barnes, S. Ourselin, Locally ranked STAPLE for template based segmentation propagation, *MICCAI Workshop on Multi-Atlas Labeling and Statistical Fusion*, 2011.
- [6] O. Commowick, S.K. Warfield, A continuous STAPLE for scalar, vector and tensor images: an application to DTI analysis, *IEEE Trans. Med. Imaging* 28 (6) (2009) 838–846.
- [7] W.R. Crum, O. Camara, D.L.G. Hill, Generalized overlap measures for evaluation and validation in medical image analysis, *IEEE Trans. Med. Imaging* 25 (11) (2006) 1451–1461.
- [8] L.R. Dice, Measures of the amount of ecologic association between species, *Ecology* 26 (3) (1945) 297–302.
- [9] I. Isgum, M. Staring, A. Rutten, M. Prokop, M.A. Viergever, B. van Ginneken, Multi-atlas-based segmentation with local decision fusion—application to cardiac and aortic segmentation in CT scans, *IEEE Trans. Med. Imaging* 28 (7) (2009) 1000–1010.
- [10] R. Heckemann, J. Hajnal, P. Aljabar, D. Rueckert, A. Hammers, Automatic anatomical brain MRI segmentation combining label propagation and decision fusion, *NeuroImage* 33 (1) (2006) 115–126.
- [11] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (3) (1998) 226–239.
- [12] S. Klein, M. Staring, K. Murphy, M.A. Viergever, J.P.W. Pluim, Elastix: a toolbox for intensity based medical image registration, *IEEE Trans. Med. Imaging* 29 (1) (2010) 196–205.
- [13] S. Klein, U.A. van der Heide, I.M. Lips, M. van Vulpen, M. Staring, J.P.W. Pluim, Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information, *Med. Phys.* 35 (4) (2008) 1407–1417.
- [14] T.R. Langerak, U.A. van der Heide, A.N.T.J. Kotte, M.A. Viergever, M. van Vulpen, J.P.W. Pluim, Label fusion in atlas-based segmentation using a Selective and Iterative Method for Performance Level Estimation (SIMPLE), *IEEE Trans. Med. Imaging* 29 (12) (2010) 2000–2008.
- [15] T.R. Langerak, U.A. van der Heide, A.N.T.J. Kotte, M. van Vulpen, J.P.W. Pluim, Label fusion in multi-atlas based segmentation using weighted atlases and surface distance measures, *Int. J. Comp.-Assist. Radiol. Surg.* 6 (8) (2013) 929–936.
- [16] T.R. Langerak, U.A. van der Heide, A.N.T.J. Kotte, F.F. Berendsen, J.P.W. Pluim, Local atlas selection and performance estimation in multi-atlas based segmentation, *Proc. ISBI* (2011).
- [17] J.B.A. Maintz, M.A. Viergever, A survey of medical image registration, *Med. Image Anal.* 2 (1998) 1–36.
- [18] C.R. Maurer, R. Qi, V. Raghaven, A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions, *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (2) (2003).
- [19] E.M. van Rikxoort, I. Isgum, Y. Arzhaeva, M. Staring, S. Klein, M.A. Viergever, J.P.W. Pluim, B. van Ginneken, Adaptive local multi-atlas segmentation: application to the heart and the caudate nucleus, *Med. Image Anal.* 14 (1) (2010) 39–49.
- [20] T. Rohlfing, R. Brandt, R. Menzel, C.R. Maurer, Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains, *NeuroImage* 21 (4) (2004) 1428–1442.
- [21] T. Rohlfing, D.B. Russakoff, C.R. Maurer, Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation, *IEEE Trans. Med. Imaging* 23 (8) (2004) 983–994.
- [22] D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, D.J. Hawkes, Nonrigid registration using free-form deformations: application to breast MR images, *IEEE Trans. Med. Imaging* 18 (8) (1999) 712–721.
- [23] M.R. Sabuncu, B.T.T. Yeo, K. Van Leemput, B. Fischl, P. Golland, A generative model for image segmentation based on label fusion, *IEEE Trans. Med. Imaging* 29 (10) (2010) 1714–1729.
- [24] M. Sdika, Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote, *Med. Image Anal.* 14 (2) (2010) 219–226.
- [25] G. Villeirs, G. De Meerleer, Magnetic resonance imaging (MRI) anatomy of the prostate and application of MRI in radiotherapy planning, *Eur. J. Radiol.* 63 (2007) 361–368.
- [26] S.K. Warfield, K.H. Zou, W.M. Wells, Simultaneous Truth and Performance Level Estimation (STAPLE): an algorithm for the validation of image segmentation, *IEEE Trans. Med. Imaging* 23 (7) (2004) 903–921.
- [27] B. Zitova, J. Flusser, Image registration methods: a survey, *Image Vis. Comput.* 21 (2003) 977–1000.