Multi atlas based segmentations: a comparison between binary and probabilistic methods

Tijl A. van der Velden, 3139433
t.a.vandervelden@students.uu.nl

January 13, 2012

1 Introduction

Manually segmenting medical images is a time consuming process. Automating this process is therefore desired. This can be done by using the information from previously segmented images, atlases. Multiple methods have been proposed to perform this task.

This paper reviews different approaches and compares them with an emphasis on the relation between probabilistic and binary methods.

In this paper, the articles listed in table 1 are described and analyzed. Section 2 gives a short introduction to atlas based segmentation. In section 3 and 4 the different methods are described. In section 3 the imaged based methods are described with a subdivision between the binary and probabilistic methods. The same subdivision is made in section 4, that describes the segmentation based methods. The different methods are compared in section 5 and a conclusion can be found in section 6.

2 Multi atlas based segmentation

With multi atlas based segmentation, the data from multiple atlases is used to segment segment a new image. An atlas is an image of which the segmentation is known. In general, two phases can be distinguished during this process.

The first phase is registration. When registering an image A to an image B, image A is geometrically aligned with image B.

The second phase is label fusion. Label fusion is the process where the segmentations of the atlases are applied to the new image. It is often the case that not all atlases agree. Some label fusion rule, a method to combine the segmentations, should be taken into account to get the final registration. These methods take the quality of the atlas, an estimation of how good an atlas will perform, into account.

	Image based	Segmentation based	Global	Local	Binary	Probabilistic
Riklin-Raviv et al. [11]	X		X			X
Ashburner and Friston [1]	X		X			X
Pohl et al. [9]	X		X			X
Prastawa et al. [10]	X		X			X
Park et al. [8]	X		X			X
Shiee et al. [17]	X		X			X
Rikxoort et al. [18]	X			X	X	
Išgum et al. [6]	X			X	X	
Sabuncu et al. [14]	X			X	X	
Commowick et al. [3]	X			X	X	
Shi et al. [16]	X			X		X
Xue et al. [20]	X			X		X
Langerak et al. [7]		X	X		X	
Warfield et al. [19]		X	X		X	
Rohlfing et al. [12]		X	X		X	
Heckemann et al. [5]		X		X	X	
Sdika [15]		X		X	X	

Table 1: Overview of articles discussed in this paper

These fusion rules can be very different. Some rules depend on the information in the images, such as differences in pixel intensities. Other rules depend on the segmentations of the atlases, A common method to combine segmentations is a majority vote:

$$S_i(x) = \arg\max_l \sum_c S_a(x) = l$$

, where S_i is the segmentation of the new image, S_a the segmentation of the atlas and x a voxel of the atlas. The similarity between the image and the atlas can be taken into account as well. After registering all atlases to the new image, the similarity between each atlas and the image can be measured. For example, the absolute difference can give an indication of the similarity. The vote of an atlas can be made more important if there is a relatively high similarity between the atlas an the image. This is a weighted majority vote procedure. As this method does not only depend on the segmentation of the atlas, but on the properties of the image as well, this method is an image based method. If the weight of the vote does not depend on the similarity of the images, but on agreement between the segmentations of the atlases, the method would be a segmentation based method.

The fusion rule can be global as well as local. The weighted majority vote procedure described above has a similarity measure based on the complete image. Therefore the method is a global method. The similarity measure could depend only on the difference of one voxel or on an area instead of the complete image. If that was the case it would be a local method.

The two different phases are often performed sequentially, although they can be applied at the same time as well.

The segmentations of the atlas can be binary or probabilistic. In a binary segmentation, a voxel is part of a particular class or not. In a probabilistic segmentation, a voxel is part of a class with some probability. A probabilistic atlas can be seen as an atlas that tries to capture the anatomical variability of a population. To get a final segmentation with a probabilistic method, a threshold of 0.5 is often applied or the class with the highest probability is selected.

3 Image based methods

3.1 Global methods

3.1.1 Probabilistic methods

Riklin-Raviv et al. propose a method in [11] where a *latent atlas* is built based on a set of images of which at most one is manually segmented. The latent atlas can segment the complete set of images. By using only a single segmented image, the time consuming process of manual delineating a set of images that covers the variance of the population would not be needed anymore.

The model consist of the images I with corresponding segmentations Γ. At the beginning only one of these segmentations is known. In a segmentation Γ_n a voxel can be part of the background or the foreground. Each image has a parameter with a Gaussian Mixture Model, θ_{I_n}, defining the intensity distributions of the segmented structure in that image. The latent spatial model parameter, θ_Γ, models the probability a specific voxel is part of the foreground. Every image has a parameter ϕ_n which is a level-set function. If $\phi_n(x) = 0$ the voxel lays on the boundary of the foreground and background, if $\phi_n(x) > 0$ the voxel is part of the foreground, otherwise it is part of the background. With this parameter the likelihood that a voxel x is part of the segmented structure can be expressed with the function $\tilde{H}(\phi_n(x))$. The goal is to maximize the likelihood of all the images and segmentations givens all the model parameters. This maximization is done using a gradient descent optimizer.

The first step initializes all the parameters. As only one segmentation, Γ_0, is known, only ϕ_0 can be calculated. This is done by calculating the signed distance to the boundary between the foreground and background for each voxel. All the other level-set functions $\phi_{1..n}$ are initialized to ϕ_0. The spatial model parameter θ_Γ is initialized to $\tilde{H}(\phi_0 \star G)$ where G represents a Gaussian kernel. The second step calculates the intensity parameters

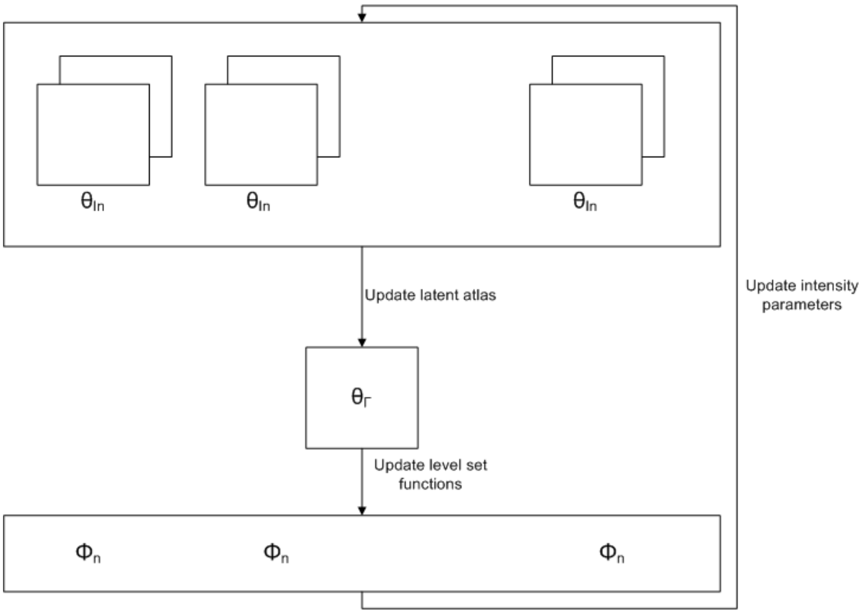

Figure 1: Schematic overview of the latent atlas approach by Riklin-Raviv et al.

of the segmented structure θ_{I_n} based on ϕ_n. In the third step the latent atlas parameters θ_Γ are recomputed based on the level-set functions as well. The fourth step updates the level-set functions by maximizing $\tilde{H}(\phi_n)$ based on θ_{I_n} and θ_Γ. The last three steps are repeated until the segmentation converges.

The method is tested by segmenting different brain structures in 39 MR images. The Dice coefficient shows similar results compared to other methods, such as FreeSurfer and an non-specified probabilistic method. Based on these results the authors suggest that a single atlas is enough to create an atlas of different images of a non-standard population. However, the authors initialized the latent atlas with multiple manual segmentations as well, which performed better than an atlas initialized with a single segmentation. This suggest that using a single atlases is not enough to create an atlas.

Although the segmentation of an image can only be part of the foreground or background, the method does not depend on the segmentations. Instead it only depends on likelihood functions based on the level-set functions. This makes that the method is a probabilistic method.

Ashburner and Friston propose a method called *Unified Segmentation* [1]. In this method, the registration and segmentation are combined into one problem. When compared to first registering followed by segmenting, this approach is computationally more expensive. However, some registration

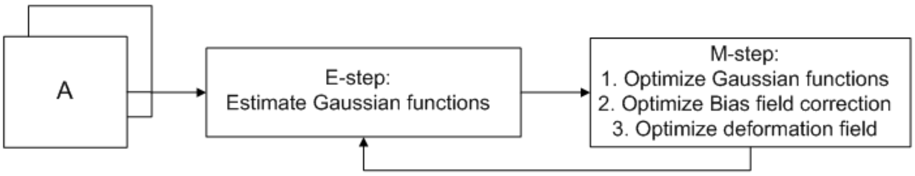
A
E-step:
Estimate Gaussian functions
M-step:
1. Optimize Gaussian functions
2. Optimize Bias field correction
3. Optimize deformation field

Figure 2: Schematic overview of the unified segmentation approach of Ashburner and Friston

parameters might be correlated to some segmentation parameters. Combining the two into one model, the complete problem is optimized instead of two optimized subproblems.

Different types of tissues are modeled with Gaussian functions. In the Gaussian functions, additional parameters are incorporated to correct for a bias field and to introduce the registration in the model. The problem is solved with an EM-algorithm. In the E-step a probabilistic atlas is estimated based on the Gaussian functions, the current transformation and the current bias field. The M-step has three consecutive steps. The first step optimizes the parameters of the Gaussian functions and the mixture coefficients of the Gaussian functions. The second step updates the bias field parameters. The last step optimizes the deformations so the registration is included in the problem.

The approach was applied to brain phantoms from the BrainWeb MR simulator with and initialized with a probabilistic atlas derived from the ICBM Tissue Probabilistic Atlas. Different types of scans were used, T1, T2 and proton density with 0%, 40% and 100% image nonuniformity (bias field). For evaluation, a threshold of 0.5 was applied on the values of the probabilistic atlas. The results, Dice similarties of more than 0.95 for the nine different phantoms, suggest that the method is very good, but it is not compared with other methods.

To get an indication of the influence of the registration and bias field parameters, the method was tested on the same images without the registration parameter, without the bias field parameter and without both parameters as well. When the registration parameter was not included in the model the results were a few percent. In the other two cases the difference became much larger, up to a 30% lower similarity.

In every iteration of the EM-algorithm a probabilistic atlas is estimated. In the first iteration, this estimation is given by a given probabilistic atlas. Therefore this method is a probabilistic method.

In [9] Pohl et al. create a Bayesian model for joint segmentation and registration. Both the segmentation and the registration parameters are updated in this method based on the Expectation-Maximization methodology. By updating the segmentation and the registration simultaneously the segmentation is not only depending on the registration, but on the differences in image intensity as well.

The method is initialized with a probabilistic atlas. This atlas is used to determine the location of the different structures. Gaussian intensity distributions are obtained from the atlas as well. A Gaussian distribution is created for each class.

In the Expectation step the probability that structure a is present at voxel x is calculated:

$$W_x(a) \triangleq P(T_x = e_a | I, \theta', R') = \frac{P(I_x | T_x = e_a, \theta'_x) \cdot P(T_x = e_a | R')}{P(I_x | \theta'_x, R')}$$

where I is the new image, T is the map, or segmentation, e_a is structure a, R are registration parameters and θ are *nuisance parameters* that represents image artifacts. The first term of the dividend is a Gaussian function based on the known Gaussian function of class e_a corrected by the nuisance parameter. The second term of the dividend incorporates the spatial parameters that are known from the probability atlas transformed by the registration parameter.

In the Maximization step the parameters R and θ are maximized,

$$R' \leftarrow \arg\max_R \sum_x \sum_a W_x(a) \log P(T_x = e_a | R) + \log P(R)$$

$$\theta' \leftarrow \arg\max_\theta \sum_x \sum_a W_x(a) \log P(T_x = e_a | \theta) + \log P(\theta)$$

.

The final segmentation is obtained by taking the structure that maximizes $W_x(a)$ for every voxel x.

The method was tested by segmenting the thalamus in 22 MR images. The thalamus is an interesting structure to segment because it has poorl visible boundaries, making it hard to register and segment correctly. Four different implementations were compared. In the first implementation the atlas was non-rigidly registered to the image and no registration parameters were used. In the second implementation was equal to the first, but used a affine transformation. The third approach used the registration parameters of the model, but only applied affine registrations. The last approach implements an hierarchical registration model. This model does not only register the images, but also depends on the structures T. When looked at the Dice coefficient, the last method outperformed the other three approaches. The results dropped however when gray matter, white matter and the ventricles are seen as a single structures, whereas they were three different structures in the initial problem. From that, the authors conclude that the method depends on easily identifiable structures.

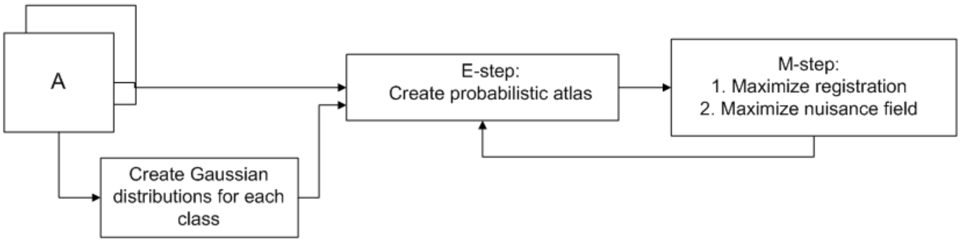
A
E-step:
Create probabilistic atlas
M-step:
1. Maximize registration
2. Maximize nuisance field
Create Gaussian distributions for each class

Figure 3: Schematic overview of the joint segmentation and registration model of Pohl et al.

In every iteration of the EM-algorithm the information of the probabilistic atlas is used. The locations of the structures in the atlas as well as the intensity distributions are taken into account. Only in the very last step of the algorithm a binary segmentation is made by assigning the class with the highest value for $W_x(a)$. Because this binary assignment is done in the last step, and all previous step make use of probabilistic values, the complete method a probabilistic method.

Prastawa et al. describe a method in [10] for automatic segmentation of developing newborn brains. Several issues make the task of automatic segmentation of the brain of a newborn more difficult than the automatic segmentation of the brain of an adult. The contrast to noise ratio is lower than that of a MR image of an adult brain due to a lower scanning resolution and a shorter scan time. This scanning time cannot be increased, as an infant cannot lay still long enough. Furthermore there are, in general, more motion artifacts. Because the white matter is myelinated during the development, the relaxation times of white matter differ inside a developing brain. Finally, the intensity ranges of the different tissues types have a large overlap.

Prior probabilities are provided by a probability atlas that is used to estimate the probability density functions (PDFs) of white matter, gray matter and cerebrospinal fluid. This atlas was created by averaging three semi-automatic segmentations. The PDFs are estimated based on voxels with a high prior probability to be part of class C_i, $\Pr(\vec{x}|C_i) > 0.9$. To distinguish myelinated and non-myelinated white matter, only samples with a low gradient are selected, so no sample is near a transition region. A minimum spanning tree (MST) of the white matter samples is created. Clusters of myelinated white matter and clusters of non-myelinated white matter are obtained by removing long edges from the MST. The need for all the prior probabilities for the different tissue types is to overcome the problem of the overlapping intensities.

Once the initial intensity distribution is known, the algorithm starts opti-

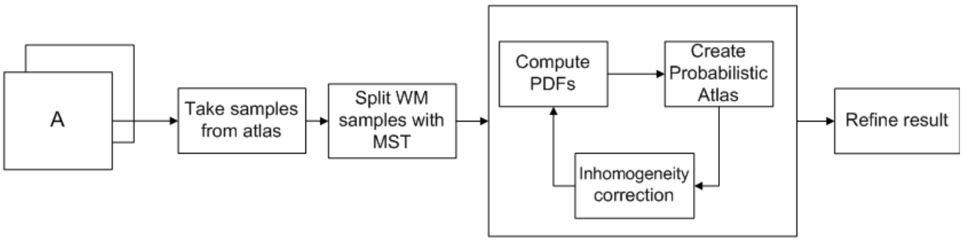

Figure 4: Schematic overview of the approach proposed by Prastawa et al.

mizing. This is done by inhomogeneity correction, where the inhomogeneity is modeled as a polynomial. With the correction, new probability distributions can be calculated and a new probability map is calculated based on the new distributions and the initial class probabilities. These steps are iterated.

In the final step, the intensity distributions are refined with a kernel density estimator and the posterior class probabilities can be calculated. A voxel is assigned to the class with the highest probability. To remove possibly outliers and false positive, the voxels are clustered in a minimum spanning tree and the outliers are pruned from the tree.

Tests were performed on four MR images from a set of 50 images. The results, quantified by Cohen's κ and the Dice similarity coefficient, show that the automatic segmentation has similar level of variability as the inter-rater variability. However, as only a small number of newborn brain MRI data was available, the probability atlas was obtained from three atlases and the method was tested on four different subjects, the method should be tested on a larger set to obtain a more representative result.

Every iteration is based on a probabilistic atlas to compute the PDFs. The result of every iteration is again a probabilistic atlas. Therefore the method is probabilistic.

Park et al. construct an *Abdominal Probabilistic Atlas* in [8]. A probabilistic atlas is used because the different organs in the abdomen have a low contrast with other organs. 32 CT images are transformed to a common coordinate space that is defined by one of the 32 atlases. The probabilistic atlas is a multi-class atlas; at each voxel the percentage of a specific class is determined by the percentage of atlases assigning that class to that voxel. The segmentation task is performed by a Bayesian framework based on Gaussian Tissue Models, the probabilistic atlas and *Markov Random Field (MRF) Regularization.*

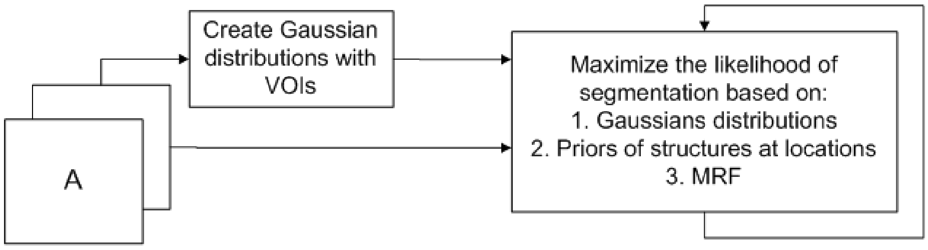
Create Gaussian distributions with VOIs
A
Maximize the likelihood of segmentation based on:
1. Gaussians distributions
2. Priors of structures at locations
3. MRF

Figure 5: Schematic overview of the approach proposed by Park et al.

The parameters of the Gaussian function of a tissue-type are estimated by selecting Voxels Of Interest (VOIs). A VOI is a voxel with a high probability, more than 0.96, to be part of a specific class. The intensities of the VOIs model the Gaussian function. Because the probability of a VOI is very high, it is unnecessary to optimize the estimations.

A penalty-function is incorporated in the form of MRF Regularization. The labels of the voxels in the six-neighborhood of voxel X_i are compared with the label assigned to X_i. For every label different to X_i, the probability that $X_i = k$ is decreased a bit.

The knowledge of the location of the structures is available in the probabilistic atlas. By using the probabilistic atlas, the probability a specific structure is present at a specific location is known as well.

The method was tested on 20 abdominal CT images. Four different structures were segmented, the liver, the right kidney, the left kidney and the spinal cord. When the Gaussian functions of the four different tissues are inspected a large overlap of the intensities is seen. This is not strange, as the structures have the same water density. When the segmentation is performed only based on the Gaussian functions and the MRF Regularization, a lot of voxels are misclassified. The two kidneys and the spinal cord are classified as the same organs, while a lot of other organs are classified as one of the four organs of interest. When the probability atlas is incorporated the four different organs were segmented a lot better with a false positive rate of 0.0054 and a false negative rate of 0.0759. Due to a large variance in the shape of the liver among people, the left lobe is often missing; the probability atlas has low values for the liver where the left lobe occurs normally.

This method is based on probability density functions as well, which are initialized with a probabilistic atlas in this case. The result is an atlas with a maximized likelihood of voxels belonging to a particular structure, in other words: a probabilistic atlas. This makes the method probabilistic.

Shiee et al. describe an *adaptive atlas* in [17]. According to Shiee et al. probabilistic atlases can distinguish structures with similar intensities quite accurate, but have a problem of segmenting subjects whose anatomy is different from the probabilistic atlas. The probability that a structure is present at a location is often used in methods using a probabilistic atlas. When the anatomy of a subject is very different from the atlas the structure in the atlas is difficult to align with the same structure in the subject. Therefore the spatial probabilities are not correct for the new subject. The goal of the adaptive atlas is to benefit from the first property but not to suffer from the second.

The approach is quite similar to the method of Park et al. [8], by estimating Gaussian parameters of structure the segmentation can be found with an EM-algorithm. The major differences are that a *Dirichlet* distribution is applied on the prior probabilities of the probabilistic atlas and in each iteration the known atlas is adapted with the new information. The adaption of the atlas changes the prior probabilities making it possible to segment structures with different shapes compared to the initial atlas.

For the first experiment the performance of the method on healthy brains was investigated and compared to a normal atlas-based EM segmentation algorithm. In the first part of the experiment a probabilistic atlas was created by averaging 18 manual delineated MR images of the brain which was applied to a brain phantom. For the second part of the experiment an atlas was created out of 8 MR images with healthy brains and applied the method to 10 subjects with healthy brains. In both cases the results of the adaptive atlas and the regular EM algorithm were very similar.

In the second experiment the performance of the method was measured on images with brains with ventriculomegaly. An atlas was created from 18 MR images with healthy subjects. The method was applied to scans of 14 different patients with hydrocephalus of which nine had moderate ventricular dilatation and five marked ventricular dilatation. The method was compared with a regular atlas-based segmentation method, with Freesurfer and with Hammer. Evaluation was quantified with the Dice coefficient and the false negative ratio. The adaptive atlas outperformed all other methods for every patient. For the patients with marked ventricular dilatation, the difference between the adaptive atlas method and the other was much larger than for patients with moderate ventricular dilatation.

Using a similar argumentation as the previous methods, this approach is seen as probabilistic. In every iteration a probabilistic atlas is used and the result is probabilistic as well, a maximized likelihood of the probability map.

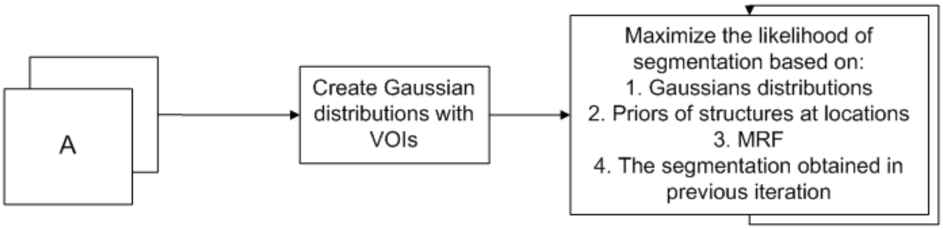
A
Create Gaussian distributions with VOIs
Maximize the likelihood of segmentation based on:
1. Gaussians distributions
2. Priors of structures at locations
3. MRF
4. The segmentation obtained in previous iteration

Figure 6: Schematic overview of the adaptive atlas proposed by Shiee et al.

3.2 Local methods

3.2.1 Binary methods

Adaptive local multi-atlas selection is a method proposed by van Rikxoort et al. [18]. The method is developed to reduce the computation time by selecting atlases. The idea of atlas selection is that poorly performing atlases can have a negative influence on the result but use computation time. By selecting an atlas based on similarity the most similar atlases are selected for a specific region. This method was applied to segment the heart and the caudate nucleus in CT scans.

The method defines regions in all the atlases. These regions can be precomputed and takes place only once. An atlas A_r is selected to be a *reference atlas*. This atlas holds the best segmentation accuracy on all the other atlases. Regions are defined in this atlas, resulting in A_{rj}, where j stands for a region. By registering A_r towards all other atlases A_i, the regions can be propagated towards the atlases, resulting in the subdivided atlases A_{ij}. The registration and propagation of the atlases creates regions the have corresponding anatomical structures.

Now the segmentation S of target image T can be computed. The first step is to define the regions in T. This is done by registering A_r towards T and propagate the regions to T and S, defining T_j and S_j.

Every S_j is updated by iterating over the following steps. In each iteration, a value p_j is calculated for each S_j which is the percentage of voxels that might change from segmentation label when a new set of propagated labels was available. If this percentage is high, there is a large disagreement between the propagated atlases and an extra atlas should be used. The locally most similar atlas $\omega(A_{ij})$ is selected, where ω is the transformation of a fast registration, and accurately registered towards T. The labels of A_{ij} are propagated towards S_j and combined with a majority vote. Then p_j is recomputed. These steps are repeated until p_j is low enough for each region.

In the final step, a single segmentation S is created by combining S_j for all j.

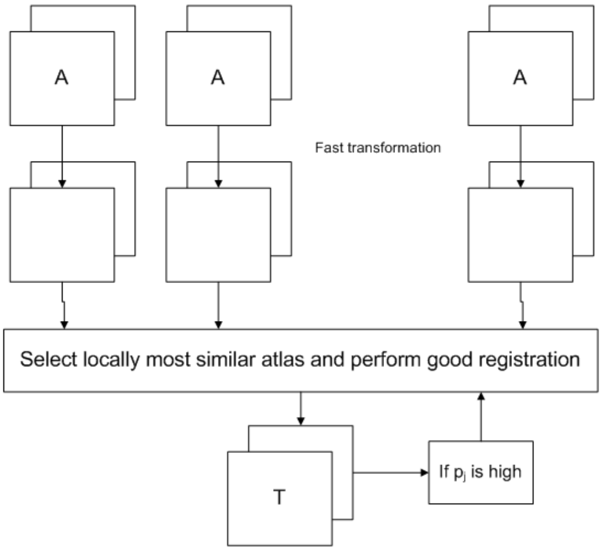
A
A
A
Fast transformation
Select locally most similar atlas and perform good registration
T
If p_j is high

Figure 7: Schematic overview of ALMAS, developed by van Rikxoort et al.

ALMAS was tested on 29 CT images of the heart and 39 MR images of the caudate nucleus. The result of *ALMAS* was similar to a majority vote procedure, where all the atlases were used. This is a positive result, as the goal was to reduce the computation time, while the performance remained the same.

All the atlases used in this method are binary. The fusion process is based on majority vote, which results in a binary segmentation. Therefore the method is one of the binary methods.

In [6], Išgum et al. propose a method named *Local Weighted Decision Fusion*. The method is based on the assumption that locally successful registrations should have more influence in the label propagation than unsuccessful registrations.

In the first step, all atlases A_i are registered towards the target image U, giving a transformation u. In the second step, for every voxel in A_i the absolute difference in image intensity D_i to the corresponding voxel in U is calculated

$$D_i(\boldsymbol{p}) = |A_i(\boldsymbol{u}(\boldsymbol{p})) - U(\boldsymbol{p})|$$

D_i is convolved with a Gaussian kernel for smoothing. The weight λ_i is calculated for each voxel in A_i, where a value of 0 means a large difference and a value of 1 means no difference in image intensity. The final

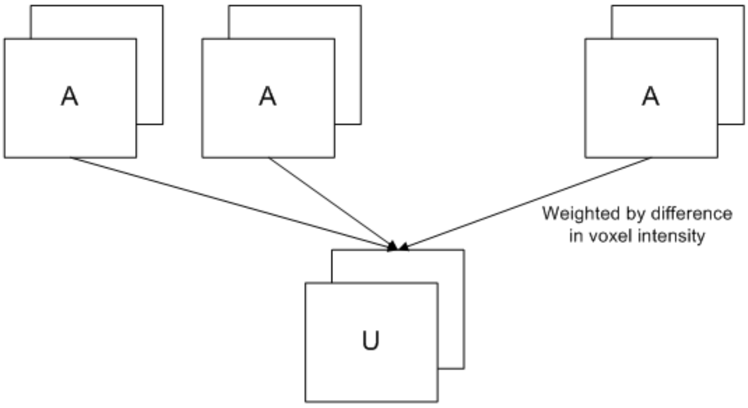

Figure 8: Schematic overview of Local Weighted Decision Fusion, Išgum et al.

segmentation S_p made by using a weighted average.

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

S_p is blurred with a Gaussian kernel and a threshold of 0.5 was used to get a binary segmentation.

The method was tested by segmenting the heart and aorta. The test-set has 29 CT images, of which 15 used as atlases and 14 as targets. Additional tests were performed on six CT images with lung diseases and four CT images with abnormalities such as metal clips and calcifications. The results, based on the Tanimoto coefficient, holds results are similar to the independent human observer results. In segmentations of the heart, no significant difference was found between the proposed method and an unweighted averaging fusion method. Not one test subject failed, but the test set was relatively small.

The method can use both binary as probabilistic atlases. The intermediate result S_p is probabilistic as well. However, this intermediate result is only used to get a more smooth final result; voxel-wise comparison is sensitive to noise. If this smoothing step was note desired, a weighted majority vote procedure would hold the same result. Because the values are not used as actual probabilistic values, this method is a binary method.

Sabuncu et al. describe a generative model for image segmentation based on label fusion [14]. Sabuncu et al. focus on label fusion methods instead of a probabilistic atlas method because it offers two advantages. Anatomical variability between subjects is better captured and multiple registrations make the method more robust for registration failures.

The method consists of a set of training images I_n with corresponding segmentations L_n. Furthermore, every voxel in the new subject $I(x)$ is obtained from at most on of the images in I_n, so the data of two different atlases are not combined into one voxel of the new subject. A random field M specifies which image that is for each voxel.

The image likelihood function $p_n(I(x); I_n)$ is a Gaussian distribution specifying the probability $I(x)$ that is generated from I_n. If the image intensity of the voxel in the new image is similar to corresponding voxel in I_n, it is likely that the voxel was obtained from atlas I_n.

The label prior $p_n(L(x) = l; L_n)$ specifies the probability the label of voxel x is l, if it was generated from label map L_n. Due to the transformations, the grid of voxels of the subject's label map L is not necessarily equal to that of L_n. To overcome this problem, this method returns the signed distance transform, normalized over all classes.

A membership prior function is defined as well. In this function, a spatial relationship is incorporated. If the six neighbors of a voxel in an image are obtained from the same image I_n, which is specified in $M(x)$, M is more likely to be correct. The influence of this function is depending on a parameter β. If $\beta = 0$ it results in a locally weighted voting procedure. If $\beta = \infty$, the method will be globally weighted. If β gets some finite, positive value, the method weighs locally similar images more. An EM-algorithm is created in the last two cases. In the E-Step a new random field M^i is estimated based on the image likelihood function, the current labelmap and the random field M^{i-1}. In the M-step a new labelmap is calculated based on the newly estimated random field M^i.

The method was tested in several ways. The first set of tests was performed on 39 brain MR images. Six subjects probably had Alzheimer's disease, five possibly had this disease and 28 were healthy. With the Dice similarity coefficient the results of segmenting different brain structures were quantified. In a leave-one-out routine, different label fusion methods were compared: FreeSurfer, Majority Voting, STAPLE, Majority10 (majority vote of the 10 most similar images), global weighted fusion ($\beta = \infty$, local weighted fusion ($\beta = 0$) and semi-local weighted fusion (β has some positive value).

The semi-local weighted fusion resulted in significant more accurate results in eight of the nine regions of interest. In the cortex, the FreeSurfer approach achieved a better result. The FreeSurfer package uses anatomical information, which is not available in the fusion methods. In all the regions, semi-local weighted fusion achieved better segmentations than local weighted fusion, which achieved better segmentation than global weighted fusion. The average difference in Dice similarity between local weighted fusion and semi-local weighted fusion was significant, but very small, less than

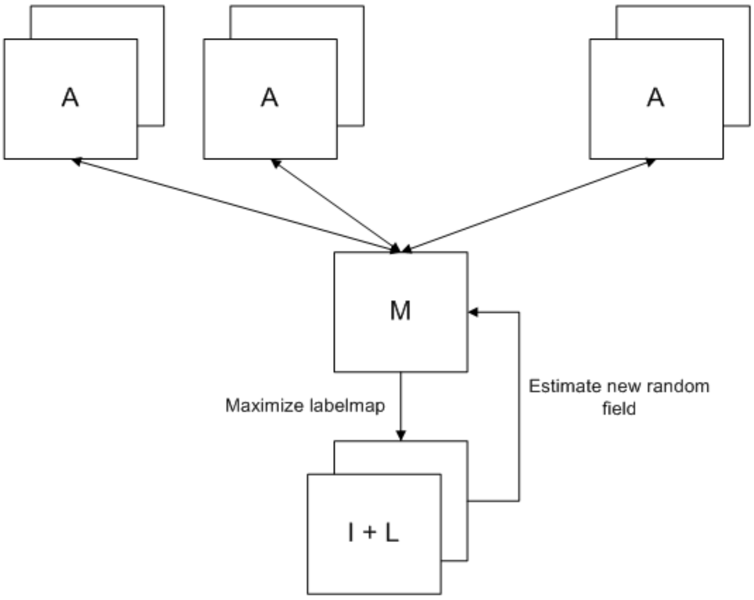

Figure 9: Schematic overview of the generative fusion model of Sabuncu et al.

0.014 for each region. STAPLE performed significantly worse than the three label fusion methods.

In a second test the volume of the hippocampus was measured in 282 MR images. The images were split up into five groups: young, middle-aged and old healthy people, people with possible Alzheimer's disease and people with probable Alzheimer's disease. The measured changes of the hippocampal volume are similar to known volume changes from medical studies. This indicates that the method segment the hippocampus in a similar way for different volumes. However, the segmentations were not compared to golden standard segmentations, as they were not available. It might be possible the method segments the hippocampus incorrect.For example, if the method oversegments the hippocampus in the first stage as well as in the second stage, the volume difference can be correct. However, the hippocampus itself is oversegmented.

The values of the atlases and labelmap can only contain discrete values of the labels. The likelihood of a correct labelmap is maximized, but only labelmaps with discrete values are taken into account. Therefore this method is binary.

In [3], Commowick et al. propose a method, based on Frankenstein's Creature Paradigm, where the most similar atlas is selected for every region.

The first goal of the method is to create an atlas that is somewhat similar to the patient, which is specifically interesting for areas with large anatomical variabilities within the population. If an atlas is very different from a patient, the registration is more likely to fail, so having an atlas that is similar to the patient increases the posibility of a successful registration. The second goal is to create a method that is not very expensive with respect to computational time.

An average atlas M is created from the atlases in the database. Regions of interest R_l are defined in M and non-linear transformations $T_{I_j \leftarrow M}$ from the atlas images I_j to M are computed. When a new image P has to be segmented, P is registered on M resulting in a transformation $T_{P \leftarrow M}$. To reduce computation time, the assumption that $T_{I_j \leftarrow P} \approx T_{I_j \leftarrow M} \circ T_{P \leftarrow M}^{-1}$ is made so the individual atlas images will not have to be registered towards the target image P. The next step is to select the most similar atlas $\tilde{I}_l$ from I_j for every region in R_l. The similarity is based on the Log-Euclidean distance on diffeomorphisms between the identity transformation and $T_{I_j \leftarrow M} \circ T_{P \leftarrow M}^{-1}$. This distance represents the magnitude of a transformation.

The following step is to create a most similar atlas, which is an iterative process. The images $\tilde{I}_l$ are registered to the average image of iteration k, $\tilde{M}_k$. A new average image M_{k+1} is computed with respect to the regions $R_{l,k}$

$$M_{k+1}(x) = \sum_{l=1}^{L} \bar{w}_{l,k}(x)(\tilde{I}_l \circ T_{\tilde{I}_l \leftarrow \tilde{M}_k})(x)$$

where T is a non-linear transformation and $\bar{w}_{l,k}(x)$ the normalized minimal distance from x to region l at iteration k. A weighted average transformation T_k is calculated from all transformations $T_{\tilde{I}_l \leftarrow \tilde{M}_k}$ with $\bar{w}_{l,k}(x)$ used as the weighing factor. T_k is applied to M_{k+1} to get $\tilde{M}_k + 1$ as well as to $R_{l,k}$ to get $R_{l,k+1}$. This is done to remain in the frame of reference of P.

The method was tested on 58 CT images of the neck and head area on structures with a high anatomical variance in the population. The specificity increased compared to a non-specified classical atlas-based method. On the other hand, the sensitivity is much lower, but this could be due to large intra- and inter-expert segmentation variability.

This method does not use the information of the segmentations of the atlases at all. Only transformations are applied to the atlas. Although the atlases can be probabilistic, the method does not use this information. Therefore this method is binary.

3.2.2 Probabilistic methods

A local, probabilistic method based on the image properties is the *multi-region-multi-reference atlas* proposed by Shi et al. [16]. The *multi-region-multi-reference atlas* tries to create a subject-specific atlas that matches the

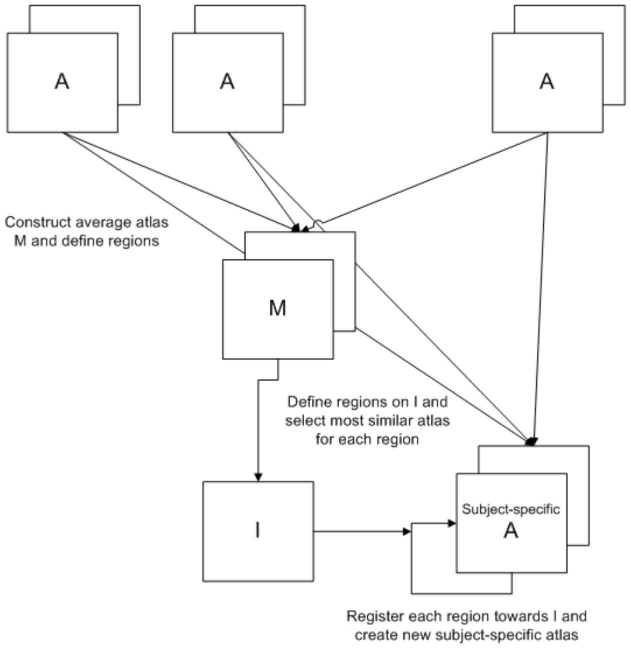
A
A
A
Construct average atlas M and define regions
M
Define regions on I and select most similar atlas for each region
I
Subject-specific A
Register each region towards I and create new subject-specific atlas

Figure 10: Schematic overview of the Patient Specific Atlas proposed by Commowick et al.

local anatomical structures to a great extent. This is done by precomputing multiple atlases for every region. The rationale behind the method is that the atlases used for region a might perform worse for region b, so other atlases have to be used to segment region b, and that multiple atlases perform better than only one.

In the first step an average image or template is constructed from all the atlases. Then all the atlases are registered to this template and a new template is constructed. By alternating between constructing the template and registering the atlases, an average shape atlas is constructed and all atlases will be warped into a common coordinate space.

The second step defines the different regions of the image. This is done with the watershed algorithm in the template. As all the atlases are in same coordinate space, the regions can be propagated to the atlases without any transformation.

The next step creates a *regional sub-population probability atlas*. Regional similar atlases are clustered together and combined into a probability atlas by calculating the weighted average. The weighing factor is a value called the representativeness that defines the likelihood that the atlas belongs to the cluster.

Then a *subject-specific atlas* is generated. The subject image is transformed to the template and the regions are propagated on the subject image. For every region the best matching sub-population atlas is is selected to be a part of the subject-specific atlas. The selection is based on the mutual information between the atlas and the new image.

In the final step a joint registration-segmentation method is used to segment the subject image. This method is based the intensity distributions of the different tissues and modeled with mixtures of Gaussians.

The method is applied to 10 MR images of neonatal brains randomly selected out of a set of 400. Eight slices were segmented in each image to create a golden standard. The result was compared to two average based atlas methods with the Dice similarity coefficient. The method outperformed the two averaged based atlases in practically every case. It is worth noting that when the proposed method performed less accurate, the average based atlases had difficulties as well.

The input atlases can be both binary as probabilistic. The multi-region-multi-reference atlas is, because of the weighted averaging, a probabilistic atlas. This probabilistic data is not used in the creation of the subject specific atlas, but it is used in the final step, the joint registration-segmentation algorithm. Therefore this method is a probabilistic method.

In [20] Xue et al. use a similar approach as Park et al. [8] to segment the cortical region in a MR image of a newborn brain. The main difference can be found in two additional steps.

The first additional step is the removal of so-called *mislabeled partial*

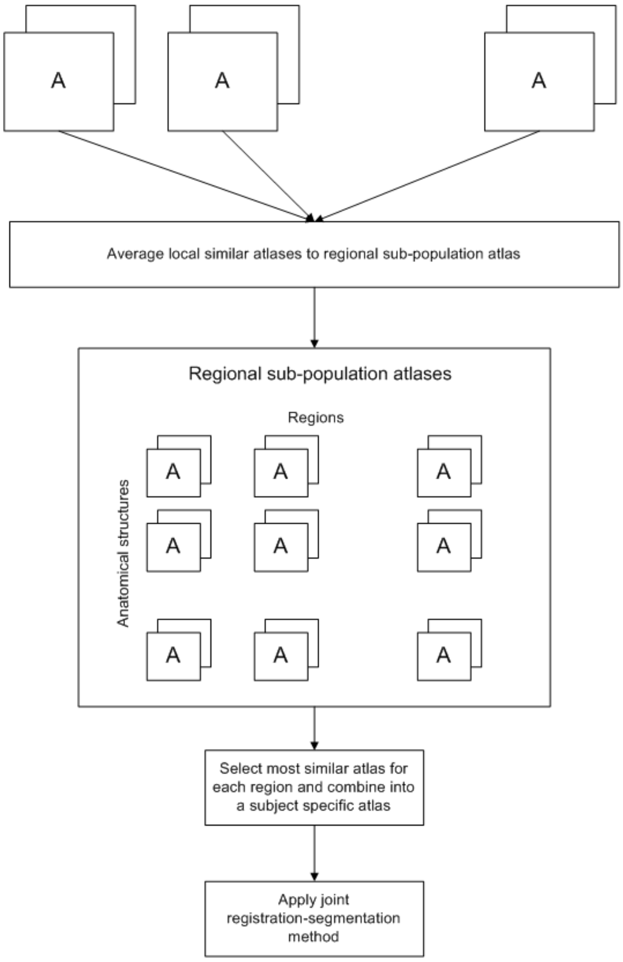
A
A
A
Average local similar atlases to regional sub-population atlas
Regional sub-population atlases
Regions
Anatomical structures
A
A
A
A
A
A
A
A
A
Select most similar atlas for each region and combine into a subject specific atlas
Apply joint registration-segmentation method

Figure 11: Schematic overview of the multi-region-multi-reference atlas of Shi et al.

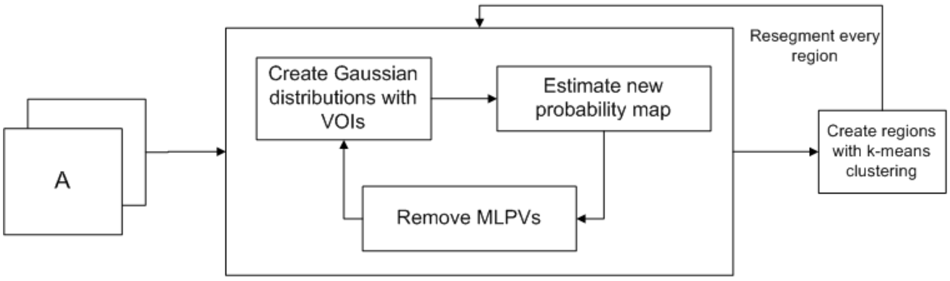
A
Create Gaussian distributions with VOIs
Estimate new probability map
Remove MLPVs
Resegment every region
Create regions with k-means clustering

Figure 12: Schematic overview of the MLPV removal approach of Xue et al.

volume voxels (MLPVs). MLPVs are voxels that partially belong to one structure, partially to another and get the label of a third structure. For example, the voxels between the CSF and the gray matter can get the label white matter in a T2-weighted image. The removal of the MLVPs is achieved by changing the influence of the Markov Random Field in each iteration.

The E-step estimates a new probability map based on Gaussian intensity distribution of the different structures. After the E-step the MLPVs are removed from the probability map with the Markov Random Field. New intensity distributions are calculated in the M-step based on the new probability map.

The second additional step takes place after the initial segmentation, where the brain is divided in sections with a k-means clustering algorithm. The clustering is performed on four parameters; the x-, y- and z-coordinate and the voxel intensity. For each section the same EM-algorithm, including the MLPV removal step, is applied again. This step is performed to overcome the problem of varying intensities of the white matter.

The method was tested on 24 different brains varying in complexity. The addition of the MLPV-step and local segmentation step show an increase of the Dice similarity coefficient. The MLPV-step removes most of the white matter voxels between the gray matter and the CSF. The local segmentation step corrects for over- and under-segmentations complex folded cortical regions.

New intensity distributions are calculated from a probability map in every E-step of the algorithm. Every M-step calculates a new probability map. Because the probabilistic data is used in every iteration, this method is probabilistic.

4 Segmentation based methods

4.1 Global methods

4.1.1 Binary methods

To get an estimation of multiple segmentations, either human raters or automated segmentation algorithms, Warfield et al. suggest a method named *STAPLE* in [19]. The original purpose of the method was to create a golden standard segmentation from multiple manual segmentations.

STAPLE is an expectation-maximization algorithm based on performance parameters; the sensitivity and the specificity of the known segmentations compared to the 'true' segmentation. In the estimation step, a 'true' segmentation is estimated based on the atlas segmentations and the performance parameters of those atlases. As the goal of the method is to find this 'true' segmentation the correct values of these performance parameters are unknown. Therefore, the sensitivity and specificity are initialized all to the same value.

In the E-step of the algorithm a segmentation is estimated. For each voxel of the image the probability that it is part of a segmentation is denoted by W_i^k where i is the voxel and k the iteration of the algorithm.

$$W_i^k \equiv \frac{a_i^k}{a_i^k + b_i^k}, \quad a_i^k \equiv \prod_{k:D_{ij}=1} p_j^k \prod_{k:D_{ij}=0} 1 - p_j^k, \quad b_i^k \equiv \prod_{k:D_{ij}=0} q_j^k \prod_{k:D_{ij}=1} 1 - q_j^k$$

p_j denotes the sensitivity of atlas j and q_j its specificity. a_i^k holds the probability that voxel i is part of the segmentation according to all atlases and b_i^k holds the probability that voxel i is not part of the segmentation. D_{ij} is the segmentation of voxel i of atlas j.

In the maximization step the performance parameters p and q of each atlas are updated based on the values of D, the manual segmentations, and W, the estimation that a voxel is part of the segmentation. These two steps are iterated until the 'true' segmentation converges. This is the case when the sum of all W_i has changed very little.

The method was tested in several ways. True segmentations of brain phantoms were available and were correctly estimated by the method and it performed better than majority vote. Some of the raters had a prespecified sensitivity and specificity. These numbers were also correctly estimated by *STAPLE*.

Although W_i is a probability map, the method itself is binary. The maximization step, where the sensitivity and specificity are calculated, uses the binary atlases D_{ij}. This is not strange, as the method is developed the combine multiple manual delineations. These delineations are, in general, binary.

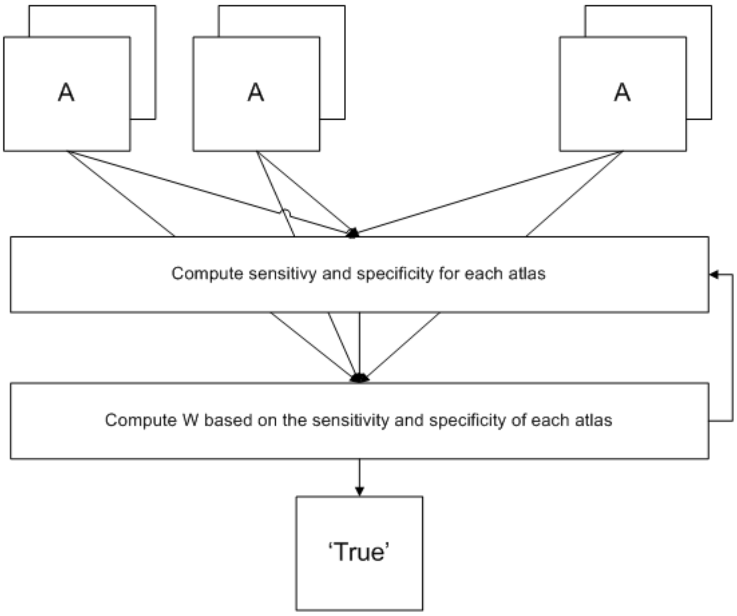
A
A
A
Compute sensitivy and specificity for each atlas
Compute W based on the sensitivity and specificity of each atlas
'True'

Figure 13: Schematic overview of STAPLE by Warfield et al.

A similar approach is suggested by Rohlfing et al. in [12]. The major difference with the method of Warffield et al. is that the method of Rohlfing et al. assumes that the different classifiers, the atlases, are independent from eachother. This is seen in the way how W is calculated:

$$W_i(x) \equiv P(x \in C_i|\boldsymbol{e}, \boldsymbol{N}) = \frac{P(x \in C_i|\boldsymbol{N}) \prod_k \lambda_{k,j,e_k(x)}}{\sum_j P(x \in C_j|\boldsymbol{N}) \prod_k \lambda_{k,j,e_k(x)}}$$

. In this equation $\boldsymbol{e}$ are the different atlases and $e_k(x)$ the label of voxel x according to atlas k. $\boldsymbol{N}$ is a matrix containing the co-occurrences of the labels in the atlas k and the labels in the 'true' segmentation and $\lambda_{k,j,e_k(x)}$ are the values of N normalized over the rows of the matrix.

In the first test 1200 deformations were randomly created from 20 3D confocal microscope images of bee brains. It was compared to the sum-rule and a generalized version of *STAPLE*. On average, the derivation of *STAPLE* outperformed the proposed method in these cases. The sum-rule performed worse than the proposed method. If, however, only a few and similar images were used, the proposed method performed better than both methods. Worth mentioning is that using more atlases resulted in a higher recognition rate for all the methods.

In the second test the 20 original bee brain images were segmented in a leave-one-out manner. Again, the results were compared to the sum-rule

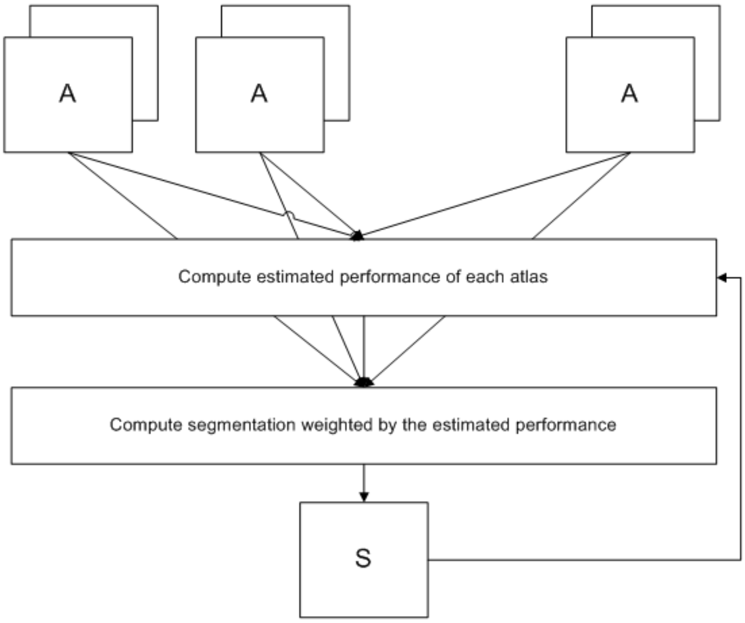
A
A
A
Compute estimated performance of each atlas
Compute segmentation weighted by the estimated performance
S

Figure 14: Schematic overview of SIMPLE by Langerak et al.

and the generalized version of *STAPLE*. In this case the proposed method performed better than the other two.

As this approach is similar to *STAPLE*, the same argumentation can be used to see this approach as a binary approach.

Langerak et al. describe a method in [7] called *SIMPLE* where poorly performing segmentations are discarded iteratively. This is an atlas selection strategy based on the assumption that poorly performing segmentations introduce noise to the final segmentation and therefore should not be used. The method is developed to combine atlas selection with performance estimation.

In the first step all the atlas images are registered towards the target image and estimation of the segmentation is made with a weighted majority vote procedure. Then, for every atlas the estimated performance ϕ_i is computed using a binary overlap measure between the atlas $L_i^{'}$ and the estimated segmentation L_{est}^k. If ϕ_i is larger than some threshold, $L_i^{'}$ will be in the set $\tilde{L}^k$. A new estimation of the segmentation, L_{est}^{k+1}, is computed by fusing the labels with a weighted majority voted procedure, weighted by ϕ_i. These steps are repeated as long as $\tilde{L}^k \neq \tilde{L}^{k-1}$.

Experiments were performed by segmenting the prostate in 100 MR images with a leave-one-out routine. The results were compared to *STAPLE*, weighted majority vote and unweighted majority vote. *SIMPLE* outperformed all the other methods. However, the method does fail occasionally. This is the case when the initial estimation, L^0_{est}, is a poor segmentation.

In every iteration the performance of each atlas is measured with a binary performance measure. This requires binary segmented atlases and a binary result. Therefore the method is binary.

4.2 Local methods

4.2.1 Binary methods

In [5] Heckemann et al. compare the quality of brain MR image segmentations of multi classifier decision fusion to segmentations based on a single atlas. The method that is used applies a very common fusion rule for binary segmentation problems: *majority vote*. With majority vote, all the atlases are registered to the image. The class with the highest agreement among all atlases for a specific voxel will be assigned to the voxel of the image.

The method was tested on 30 MR images of the brain with a leave-one-out routine. A total of 67 different structures were segmented in these images and used for the quantitative analyses. Twelve of those structures were closely examined to see the effect of differences in location and shape. The performance was measured with the Dice similarity coefficient and compared to two single atlas based methods. The effect of the number of atlases used was measured as well and varied from 3 to 29.

An increase of similarity was obtained when more atlases were used. Compared to single atlas based methods, using the majority vote rule resulted in better segmentations as well. In general the results were quite accurate, especially for larger structures. More complex structures, such as the temporal horn and several gyri, were much harder to segment correctly.

A majority vote procedure can only use binary atlases. The result of such a procedure is binary as well, the label with the most votes will be assigned to a voxel. This makes the method binary.

Sdika proposes a method for a weighted majority vote in [15]. The methods is based on the idea that atlas that performs well in most cases should have more weight in the voting procedure. For every atlas an accuracy map q is created. This map holds the accuracy for each voxel, where the accuracy is defined as the percentage of other atlases that, after transformation T, would apply the same label at that voxel.

$$q_j(x) = \frac{1}{N_a - 1} \sum_{i \neq j} T_i^{-1}(g_i(x))$$

q_j is the accuracy map of atlas j. $g_i(x)$ is 1 if, after registration of image i on j, the labels of the voxels in i and j are the same. These computations are performed only once.

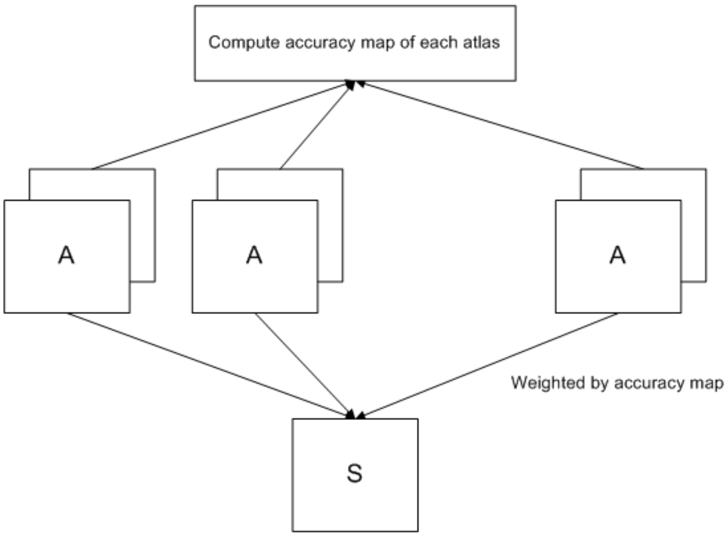

Figure 15: Schematic overview of approach using accuracy maps by Sdika.

The segmentation of the new subject is created with an *accuracy weighted vote*. This is a weighted majority voting procedure.

$$L(x) = \arg\max_{l} \sum_{L_n(T_n(x))=l} q_n(x)$$

where T_n is a transformation from I_n to the new image I.

The method is tested on 18 MR images of the brain from the IBSR data set. The results are compared with a normal majority voting procedure. The weighted voting procedure outperforms the normal voting procedure in all the tests performed.

This method only uses binary atlases. The result of a weighted majority vote procedure is binary as well. This makes the whole method binary.

5 Analysis

Heckemann et al. and Sdika et al. showed that combining the information of atlases holds more information than using the information of just one. Sdika also changed the influence of an atlas based on a performance measure. This measure was created on the atlas-set itself. For most cases this results in a better segmentation. If, however, the new subject is anatomically different from the training-set, the algorithm is likely to fail.

To overcome the problem where methods fail to segment different subjects several alternatives have been proposed.

Commowick et al. and Išgum et al. try to improve the result by looking at local image similarity. Van Rikxoort et al. uses image similarity to make the process more efficient. The method of Commowick et al. selects the most similar atlas for each region. Large anatomical variances are covered with this approach.Whereas Commowick et al. use one atlas per region, van Rikxoort et al. use multiple atlases for each region and showed that using multiple atlases for each region has a positive influence on the result. Išgum et al. look at voxel-wise similarity instead of regional similarity and obtain similar results as van Rikxoort et al. There is no need to define regions with voxel-wise comparison, which can be seen as an advantage. However, voxel-wise comparison is much more sensitive to noise, so a trade-off has to be made for each application. What all three methods do indicate is that the local assessment of the image similarity results in better segmented images for binary methods.

Shi et al. create a probabilistic atlas based on local image similarity as well. The difference with the previous methods is that after the fusion of the atlases another step is made based on the result, a joint segmentation and registration routine. A similar routine is described by Prastawa et al. who use an averaged atlas and not a *multi-region-multi-reference atlas.* Both methods are used to segment the brain of a newborn and tested on a relatively small group. The better results achieved by Shi et al. suggest that selecting atlases on local image similarity has a positive effect on joint registration and segmentation routines.

The method described by Pohl et al. is a combined registration and segmentation routine. Instead of solving the two subproblems of the process, a single problem is created to find one optimal solution. The global approach used holds very good results, but these are hard to interpret as the success of the segmentation of the thalamus strongly depends on the segmentations of the surrounding structures. The results of these segmentations are unfortunately not given.

The results of Ashburner and Friston are difficult to interpret as well. The method was tested on a small set of fictive data. In the different experiments it became clear that the registration parameters did not add much to the result. Pohl et al. compared different variances of registration as well and found much larger differences. These findings seem contradictory, but according to Pohl et al. the assumption is made that '*the atlas is globally aligned to the image space.*' [9] in the method of Ashburner and Friston, which could explain the small differences.

Park et al. segment an abdominal image based on voxel intensity and prior probability.The probabilistic atlas used for the prior probabilities is an averaged atlas. Therefore some anatomical differences have a low prior probability and are not found by the method of Park et al. This problem is solved mostly by Xue et al. who, after a global segmentation, segment different regions individually. If, however, the prior probability of a structure at that location is too low, this method will not help much. The method suggested by Shiee et al. to solve this problem seems very promising. But it should be taken into account that the intensity of the changed structure, the ventricles, has a large contrast with the intensity of the gray matter, which is not the case for the organs in the abdomen. The multi-region-multi-reference atlas of Shi et al. could be applicable in this case as well. However, atlases with the specifc anatomical difference have to be available.

Riklin-Raviv et al. use only one manual segmented image to segment an ensemble of images. The results achieved seem very promising. However, Blezek et al. showed in [2] that a population is often not described best by only one atlas and Išgum et al. showed improved results when more atlases are used (up to an optimum). Riklin-Raviv et al. showed this as well, but the differences between the results were much smaller. This means that if multiple atlases are available it is wise to use all of them instead of using only one. The method might be more useful for intra-subject segmentation, such as the tumor segmentation task described by Riklin-Raviv et al. as the area to segment is similar in all the images.

The method described by Warfield et al., *STAPLE*, and its derivation described by Rohlfing et al. assume that the segmentations are all made in a similar way. '*Implicit in this model is the notion that the experts have been trained to interpret the images in a similar way, the segmentation decisions may differ due to random or systematic rater differences, ...* [19]' An incorrect registered or anatomically very different image can be seen as an expert interpreting the image in a different way and therefore corrupting the method. Or as Crum et al. say in [4] '*It is not clear how far these assumptions will apply to automated segmentation techniques.*' *SIMPLE* does not make such an assumption and will ignore inaccurate atlases completely. '*The main difference with STAPLE is that in each iteration badly performing segmentations are discarded. These segmentations no longer contribute to the estimate of the ground truth segmentation*' [7]. The better results achieved by *SIMPLE* suggest that the assumption made by Warfield et al. is not applicable to multi atlas based segmentation problems.

When we look at the probabilistic approaches it can be seen that all the approaches use the probabilistic atlas to create Gaussian functions for an intensity-based segmentation approach in areas with low contrast between the different structures. According to Prastawa et al. '*the use of a probabilistic atlas [...] is crucial to overcome the intensity contrast limitations.*' [10], which explains the choice of using of such an atlas in all these approaches. The Gaussian functions are probability density functions in these cases. These functions are modelled more accurate with probabilistic values than a discrete segmentation. However, it is important that the atlas is somewhat similar to the new subject. Most approaches create an atlas by averaging multiple manual delineations. Some anatomical variances might be removed by this approach. Shi et al. reduce this problem by creating a patient specific atlas and Shiee et al. by making the atlas adaptive. In the case of Shi et al. the variance of anatomy has to be available in on of the atlases. For the method of Shiee et al. it is the question if it works for low contrast regions as good as it did for high contrast regions.

The binary approaches combine the atlases directly into a result. All the atlases are used and registered to the new subject. This is computationally expensive, but holds better results than an averaged atlas [2] that are used in probabilistic methods. According to Sabuncu et al. the main advantages of multi atlas binary approach are '*(1) across-subject anatomical variability is better captured than in a single atlas, which can be viewed as a parametric model that typically uses single mode distributions (e.g., Gaussian) to encode anatomical appearance, and (2) multiple registrations improve robustness against occasional registration failures.*' [14]. Similar argumentations are presented in [13] and [5].

The argument of Prastawa et al. to use a probabilistic atlas might be the reason that no segmentation based method is found using probabilistic atlases. As the segmentation based methods do not use any information from the image, the intensity contrast limitations cannot be a problem for these approaches. Prior knowledge of a structure at a location is useful, but this knowledge is often added in a different way. In the method of Langerak et al. a method known to perform acceptable is applied to obtain this prior knowledge. For the method of Warfield et al. this knowledge is incorporated in the assumption that all experts interpret an image in a similar way.

6 Conclusion

In this paper several atlas based segmentation methods are discussed with the emphasis on the differences between binary and probabilistic approaches.

A probabilistic atlas is most useful when a relation between the intensity of a voxel and an anatomical structure has to be found, which is the case with intensity based classifiers. Often only classifying based on the intensity is not sufficient and prior probabilities of the availability of structures on a specific location are incorporated as well. As the atlas is often an average of multiple atlases, rare anatomical differences are hard to segment with these methods, especially when different structures have the same intensity range.

Several ideas have been suggested to overcome this problem. The creation of a patient specific atlas creates an atlas more similar to the new subject. This requires the availability of the anatomical variance in the set of available atlases. An adaptive atlas can be used as well. This does not require a similar atlas, but it is still unknown how well this approach works with changed structures that have a small intensity contrast with neighboring structures.

The advantage of probabilistic methods, compared to most binary methods, is that they require a relatively small amount of computational time. In the registration-segmentation problem the registration step is often the most expensive step from a computational point of view. A patient has to be registered to only one atlas and the atlas is created only once.

The binary methods are computationally much less efficient. If no atlas selection is applied before registration, every atlas is registered towards the new subject. But often this computational burden is not a real problem, especially when no real time result is required. The positive side of the binary methods is the ability to segment anatomically rare variances of structures more easily. In contrast to probabilistic atlas, where these variances are averaged away, the information of variances remains available in binary methods. Using all these atlases is most effective when some kind of similarity measure is used in the fusion process. A local similarity measure is in general better than global similarity measure. Voxelwise similarity measures are fast but sensitive to noise. When neighboring voxels are taken into account the noise has less influence but the method requires more computational power.

Another positive effect of multiple registrations is the robustness against registration failures. If a subject is registered to a single atlas, which is the case with probabilistic methods, the method will fail when the registration fails. With binary multi atlas methods, a small subset of the available atlases can fail in the registration step but the overall performances might be very acceptable.

Several articles showed that a binary method might not need all the available atlases; using a subset of similar atlases resulted in good segmentations as well. Selecting atlases prior to registration can reduce the computational costs and should be considered.

References

The key references are marked with a * in front of it.

[1] John Ashburner and Karl J. Friston. Unified segmentation. *NeuroImage*, 26(3):839 – 851, 2005.

[2] Daniel J. Blezek and James V. Miller. Atlas stratification. *Medical Image Analysis*, 11(5):443 – 457, 2007.

[3] * Olivier Commowick, Simon Warfield, and Grégoire Malandain. Using frankenstein's creature paradigm to build a patient specific atlas. In Guang-Zhong Yang, David Hawkes, Daniel Rueckert, Alison Noble, and Chris Taylor, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, volume 5762 of *Lecture Notes*

in Computer Science, pages 993–1000. Springer Berlin / Heidelberg, 2009.

[4] W.R. Crum, O. Camara, and D.L.G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *Medical Imaging, IEEE Transactions on*, 25(11):1451 –1461, nov. 2006.

[5] * Rolf A. Heckemann, Joseph V. Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115 – 126, 2006.

[6] * I. Isgum, M. Staring, A. Rutten, M. Prokop, M.A. Viergever, and B. van Ginneken. Multi-atlas-based segmentation with local decision fusion - application to cardiac and aortic segmentation in ct scans. *Medical Imaging, IEEE Transactions on*, 28(7):1000 –1010, july 2009.

[7] * T.R. Langerak, U.A. van der Heide, A.N.T.J. Kotte, M.A. Viergever, M. van Vulpen, and J.P.W. Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). *Medical Imaging, IEEE Transactions on*, 29(12):2000 –2008, dec. 2010.

[8] Hyunjin Park, P.H. Bland, and C.R. Meyer. Construction of an abdominal probabilistic atlas and its application in segmentation. *Medical Imaging, IEEE Transactions on*, 22(4):483 –492, april 2003.

[9] * Kilian M. Pohl, John Fisher, W. Eric L. Grimson, Ron Kikinis, and William M. Wells. A bayesian model for joint segmentation and registration. *NeuroImage*, 31(1):228 – 239, 2006.

[10] * Marcel Prastawa, John H. Gilmore, Weili Lin, and Guido Gerig. Automatic segmentation of mr images of the developing newborn brain. *Medical Image Analysis*, 9(5):457 – 466, 2005.

[11] * Tammy Riklin-Raviv, Koen Van Leemput, Bjoern H. Menze, William M. Wells III, and Polina Golland. Segmentation of image ensembles via latent atlases. *Medical Image Analysis*, 14(5):654 – 665, 2010.

[12] * T. Rohlfing, D.B. Russakoff, and C.R. Maurer. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *Medical Imaging, IEEE Transactions on*, 23(8):983 –994, aug. 2004.

[13] Torsten Rohlfing, Robert Brandt, Randolf Menzel, and Calvin R. Maurer Jr. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, 21(4):1428 – 1442, 2004.

[14] M.R. Sabuncu, B.T.T. Yeo, K. Van Leemput, B. Fischl, and P. Golland. A generative model for image segmentation based on label fusion. *Medical Imaging, IEEE Transactions on*, 29(10):1714 –1729, oct. 2010.

[15] Michaël Sdika. Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. *Medical Image Analysis*, 14(2):219 – 226, 2010.

[16] * Feng Shi, Pew-Thian Yap, Yong Fan, John H. Gilmore, Weili Lin, and Dinggang Shen. Construction of multi-region-multi-reference atlases for neonatal brain mri segmentation. *NeuroImage*, 51(2):684 – 693, 2010.

[17] * Navid Shiee, Pierre-Louis Bazin, Jennifer Cuzzocreo, Ari Blitz, and Dzung Pham. Segmentation of brain images using adaptive atlases with application to ventriculomegaly. In Gábor Székely and Horst Hahn, editors, *Information Processing in Medical Imaging*, volume 6801 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin / Heidelberg, 2011.

[18] * Eva M. van Rikxoort, Ivana Isgum, Yulia Arzhaeva, Marius Staring, Stefan Klein, Max A. Viergever, Josien P.W. Pluim, and Bram van Ginneken. Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus. *Medical Image Analysis*, 14(1):39 – 49, 2010.

[19] * S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *Medical Imaging, IEEE Transactions on*, 23(7):903 –921, july 2004.

[20] * Hui Xue, Latha Srinivasan, Shuzhou Jiang, Mary Rutherford, A. David Edwards, Daniel Rueckert, and Joseph V. Hajnal. Automatic segmentation and reconstruction of the cortex from neonatal mri. *NeuroImage*, 38(3):461 – 477, 2007.