SAHLGRENSKA ACADEMY

ATLAS-BASED SEGMENTATION OF ULTRAHIGH-RESOLUTION STRUCTURAL MR HEAD IMAGES ACQUIRED AT 7 TESLA

Frida Johansson

Essay/Thesis:	30 hp
Program and/or course:	Medical physics
Level:	Second Cycle
Semester/year:	Autumn 2018
Supervisor:	Rolf A. Heckemann, Emilia Gryska
Examiner:	Magnus Båth

Abstract

Essay/Thesis:	30 hp
Program and/or course:	Medical physics
Level:	Second Cycle
Semester/year:	Autumn 2018
Supervisor:	Rolf A. Heckemann, Emilia Gryska
Examiner:	Magnus Båth
Keyword:	Anatomical segmentation, Brain, MRI, 7 Tesla, Pincram, MAPER, Shape based averaging, Ultrahigh resolution

---

Purpose: The purpose of this work was to find out how the existing brain atlases and segmentation algorithms perform when applied to ultrahigh-resolution MR brain images, acquired with a 7-Tesla scanner. Also to make adaptations to deal with the potential challenges and evaluate the quality of the anatomical segmentations of the 7-Tesla images.

Materials: A dataset of MR brain images with various resolutions (1mm, 500 μm, 250 μm non-averaged & 250 μm averaged) shared by Lüsebrink et al. from the 7 Tesla scanner in Magdeburg was used.

Methods: Two atlas-based anatomical image segmentation algorithms were applied: Pincram for brain extraction and MAPER for labelling multiple brain regions. The resulting brain masks and label maps were assessed qualitatively and quantitatively. Visual evaluation of the quality of the segmentations was made by the auther and external experts. To quantify the consistency of segmentations at the highest resolution, the Jaccard overlap coefficient were calculated.

Shape base averaging (SBA) has been implemented on the MAPER-segmented atlases and applied to a 500 μm resolution image to improve the appearance of the segmentation. It was then compared to Vote Rule decision Fusion (VRF) that is the standard method of fusing atlas labels in MAPER.

Conclusion: MAPER and Pincram work on brain images obtained with a 7-Tesla scanner even though the algorithms have been designed for and validated on 1.5 and 3 Tesla. The data size at the highest resolution exceed available computational resources, therefore images had to be downsampled to 500 μm.

The segment boundaries were smoother with SBA than with VRF and they got more pleasant to look at. Some boundaries do get misplaced, so the volume estimation of the structures might not be better than with VRF.

Table of content

Introduction

Magnetic resonance imaging (MRI) is a widely used modality for structural brain imaging due to its great contrast between different brain tissues and its high spatial resolution. MR imaging can be carried out non-invasively and, unlike X-ray computed tomography (CT), does not expose the study subject to ionizing radiation. Scanners commonly used for clinical purposes operate with field strengths of 1.5 or 3 Tesla, but recently, 7-Tesla scanners have been introduced. With these, it is possible to acquire MR images with ultrahigh resolution. The advantages of having ultrahigh resolution are improved visualization of small details, higher signal to noise ratio (SNR), and diminished partial volume effect. Such images are, however, more sensitive to artefacts, for example artefacts caused by movements [1]. The majority of the scanners with higher field strength than 3 Tesla are used for research, but there is at least one 7-Tesla scanner located in North America and one in Sweden used clinically [2].

Anatomical segmentation of brain images plays an important role in clinical research. It is used to determine the volume of the whole brain as well as its anatomical structures. Volumetric analysis of the segmented regions can be used to support diagnosis and track progression of neurodegenerative diseases such as Alzheimer's disease (AD) [3]. For example, pronounced whole-brain shrinking, ventricle enlargement, and hippocampal atrophy in particular are characteristics of AD. These changes can be detected and quantified in vivo with anatomical brain image segmentation [4].

Ultrahigh resolution will potentially lead to improved brain image segmentation. More accurate volume measurements of brain structures may entail higher sensitivity for diagnosing neurological diseases. It may also enable detection of subregional changes in the hippocampus or give rise to structural imaging biomarkers with high specificity.

Manual segmentation of brain images is currently the gold standard in clinical research. The process is time consuming and requires profound knowledge about the anatomy of the human brain. Therefore, manual segmentation is not suitable for some tasks, especially for analysing a large cohort of subjects. For this, the segmentation process needs to be automated. One of the challenges in automatic anatomical segmentation is that many anatomical structures of the brain consist of more than one tissue type, leading to a mixture of signal intensities within a region in the MR image. Therefore conventional image segmentation techniques based on intensity classification are insufficient. One way of achieving automatic anatomical segmentation is to propagate a handmade segmentation of one individual to another individual's brain by using image registration to estimate the correspondence between the anatomies of the pair [5]. This technique of doing label propagation makes up the basis of MAPER, an automatic anatomical brain image segmentation tool [6], and Pincram, a brain extraction tool [7].

It has been shown that Pincram and MAPER are robust and accurate when applied to brain images with a resolution of approximately 1 mm acquired at 1.5 or 3 Tesla [6, 7]. It is, however, not known how these algorithms perform on ultrahigh-resolution images obtained on 7-Tesla scanners. The potential benefits of anatomical segmentation of the brain on ultrahigh-resolution images thus have not yet been quantified. The scope of this thesis was to evaluate the segmentation of brain images with ultrahigh resolution. The comparison of the segmented 7-Tesla images with the segmented 3-Tesla images is left to future studies.

The aims of this thesis project were 1) to find out how the existing brain atlases and segmentation algorithms perform when applied to ultrahigh-resolution MR brain images; 2) to identify necessary adaptations to deal with the potential challenges that such images pose; and 3) to validate the results of the segmentation of ultrahigh-resolution MR brain images.

Materials and Methods

Dataset

Atlases

An atlas is defined as an intensity image (MR image) and a delineated region (or regions) of interest (label) in that image. Typically, an atlas is created by an expert who manually traces the region of interest on each image slice. Brain atlases can be used as an aid to trace brain anatomy by unexperienced observers. Atlases have also become commonly used in the domain of automated anatomical brain image segmentation. Due to the variations in shape and size of individual brains and brain structures, availability of numerous atlases improves the ability to capture inter-subject anatomical variability.

The brain atlases used to carry out the automatic anatomical segmentation of the target brain were created by Hammers et al [8-11]. The atlases were made by manually segmenting 30 MRI brain scans of 30 healthy adults. Each atlas consists of 95 segments that have been delineated by trained experts in several phases.

An example of how the manually segmented brain image can look like is shown in Figure 1. To the left is an axial section of atlas number 27 from the Hammers atlas, and to the right is a zoomed-in image of the upper right part of the image.

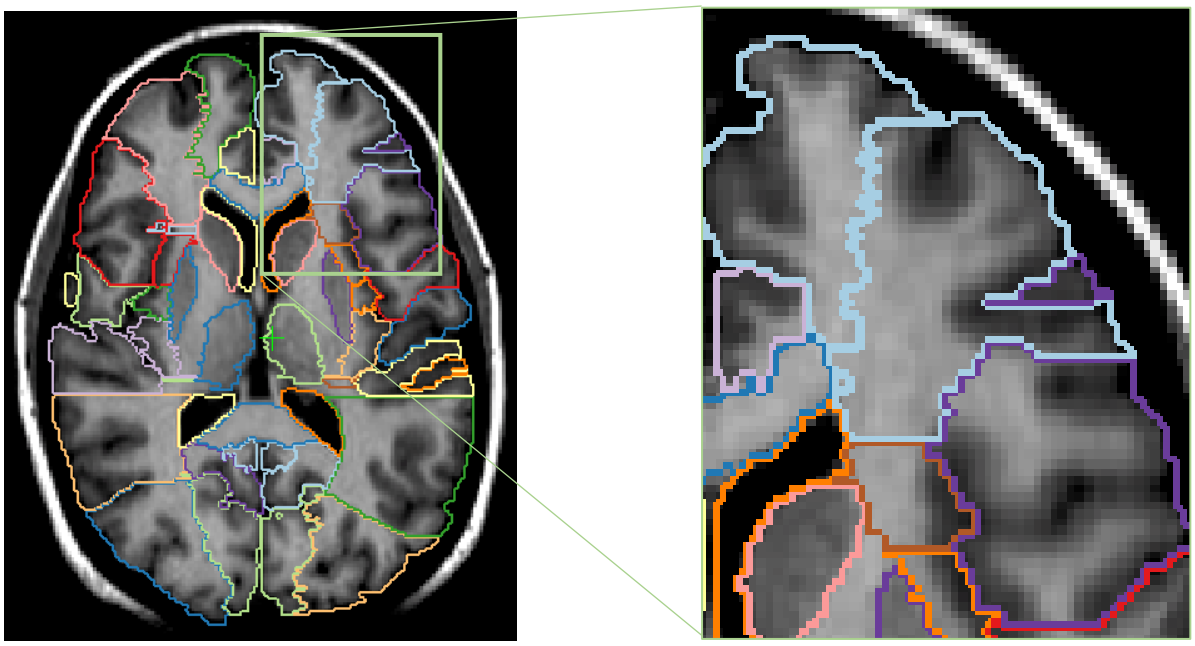


Figure 1. An axial slice of atlas number 27 from the Hammers atlas with a zoomed-in image of the top right part of the image.

7T dataset

The 7-Tesla dataset used consisted of T1-weighted whole-brain images of a single healthy subject (born 1982) acquired at the 7-Tesla facility in Magdeburg, Germany, and publicly shared by the investigators. Full technical details are available in the data paper by Lüsebrink et al. [12]. Briefly, eight ultrahigh-resolution brain volumes were acquired in five sessions with a resolution of 250 μm and averaged. In two additional sessions, images with lower resolution (1 mm and 0.5 mm, respectively) were acquired. Magnetization Prepared Rapid Acquisition Gradient Echo (MPRAGE) sequence and prospective motion correction were used. Optical Moiré phase tracking was used to accomplish prospective motion correction.

The reconstructed volumes were corrected for field inhomogeneities using a customized script for Statistical Parametric Mapping (SPM). To align the anterior commissure (AC) and posterior commissure (PC) in the same axial slice "acpcdetect" from the Automatic Registration Toolbox (ART) [13] was applied. For three of the sessions, more than one volume of the brain was acquired and averaged after rigid registration. Before averaging of the images, they were aligned with a diffeomorphic registration method from the Advanced Normalization Tools (ANTs) [14].

One notable difference between the images acquired at 7 Tesla respectively 3 Tesla with the MPRAGE-sequence is that the cross-section of the vessels appears as white dots in the T1-weighted 7-Tesla images. This is because of the lack of body RF transmit coils for the 7-Tesla scanner. When the body coil is used for the 3-Tesla scanner, the blood will be inverted far from the imaging volume (the head). In contrast, the blood does not get inverted until it reaches the head when not using a body coil in the 7-Tesla scanner [15].

Pincram

The main purpose of Pincram is to create a brain mask in the target image that distinguishes brain voxels from adjacent tissue- and background voxels. The brain mask can be used to estimate the volume of the brain, or to perform further segmentation within the brain mask. Pincram also has the ability to calculate and show the accuracy of the segmentation. It determines the "success index" by calculating the Jaccard coefficient (see below) between the final fused brain mask and the brain mask generated first in the prior iteration.

Pincram is an algorithm that propagates multiple atlases and iterates from coarse to more refined image registration to find the most accurate delineation of the target brain. In the first iteration, the brain labels are propagated by rigid registration. In the second iteration, an affine image registration is used, and the last – a non-linear registration is applied [7].

The difference between the three image registration types is the number of degrees of freedom used when registering two different images. Rigid registration has six degrees of freedom which is translation and rotation in three dimensions. In addition to the six degrees of freedom of the rigid registration, affine registration has six more degrees of freedom, being scaling and shearing in three dimensions. The degrees of freedom for the non-linear registration depends on how many control points are used and the spacing between the control points. Decreasing the space between the control points leads to more degrees of freedom [16].

The following procedure is repeated at every refinement level to generate a brain mask and select a subset of the atlases to be used in the subsequent level (a schematic overview is shown in Figure 2). First, the label sets are propagated to the target brain using the current registration type. The generated label sets are then fused in the geometrical space of the target image by summation, thresholding and binarization to create an estimated brain mask of the target brain (C_l). A margin mask (M_l) consisting of all voxels in the vicinity of the estimated brain mask boundary (up to a distance of four voxels) is created. By comparing the target image with every transformed atlas image within the margin mask, a ranking of the atlases is obtained. The atlases that are most similar to the target within the margin region are then used in the next level. To make the subsequent level more precise, a second brain mask is created by thresholding and binarizing the fuzzy label (F_l) that has been created by summing the most similar transformed labels. The voxels with values between 15% and 99% in the fuzzy label are converted to 1 and the rest of the voxels to 0 to create the second margin mask (M'_l). The second margin mask is then applied to the target to determine the parameters for the registration in the subsequent level. This ensures that the margin used for the ranking of the atlases contains the true boundary of the target brain [7].

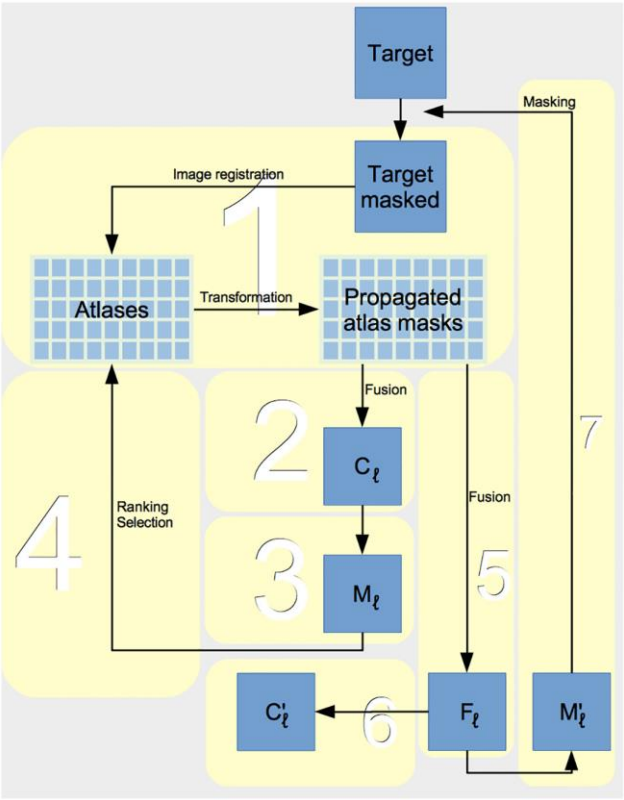


Figure 2. Flow chart of the iterative procedure of the Pincram algorithm [7].

MAPER

MAPER is a whole-brain segmentation algorithm that also propagates multiple atlases to the target brain with image registration. Before the target image can be segmented, brain extraction and bias field correction are applied as pre-processing steps. The brain extraction can be done with Pincram. Each atlas is aligned with the target image, first with rigid, then affine and finally non-linear registration creating *n* possible label maps for the target. Every voxel in the target image has *n* label values corresponding to the *n* atlas labels at that voxel. The final decision which label is assigned to a given voxel is made by vote-rule decision fusion (VRF) [5, 6] that assigns the label indicated by the majority of the atlases [17].

Jaccard coefficient

The Jaccard coefficient (JC) is a measure of the overlap between two labels (A and B). It is defined as the intersection of the labels divided by the union of the labels,

$$\mathrm{JC} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|} \tag{1}$$

and takes a value between 0 and 1. The intersection (left) and union (right) of the labels are schematically shown in Figure 3. In the ideal case where the segments completely overlap each other, the ratio becomes 1 and it becomes 0 if they do not overlap at all.

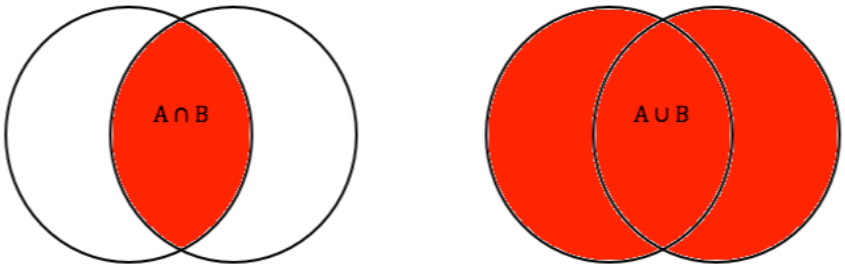


Figure 3. Left: The intersection between section A and B. Right: The union of section A and B.

Ultrahigh-resolution image

Pincram

The eight ultrahigh-resolution (250 μm) images acquired at five different sessions and the average between the eight images were downsampled to 500 μm with cubic interpolation. This was necessary because Pincram processing at the native resolution failed due to excessive runtime requirements at the shared cluster of CPUs used. To segment one 250 μm image with 30 atlas images, a full node with 32 CPU cores would have been occupied for 360 hours, which would potentially have blocked more than 900 processes by other users. Visual assessments of how well the brain masks delineate the brain were carried out for the original 500 μm resolution image and the downsampled images. The Jaccard overlap coefficients were calculated to determine the level of agreement between the 9 brain masks that were acquired with the downsampled images.

MAPER

The downsampled, ultrahigh-resolution images and their corresponding brain mask made with Pincram were used to carry out brain segmentation with MAPER. The segmentation of the brain images with ultrahigh resolution was visually compared with each other and with the segmented 500 μm resolution brain image. To get a measure of the similarity between the segmentations of all of the downsampled images, the mean Jaccard coefficient of all 95 segments were calculated between the segmentations of all the 9 downsampled images.

The 8 segmentations of the downsampled non-averaged images were fused with VRF and then applied to the averaged image with 250 μm resolution. The mean Jaccard overlap coefficients for all 95 segments were calculated between the fused segmentation and 9 MAPER segmentations acquired with the downsampled images.

Shape-based averaging

Instead of using VRF when fusing the atlases in the scripts for MAPER and Pincram, other methods can be used. One of the methods is the shape-based averaging (SBA) method that has been shown to produce more regular and contiguous labels than VRF [18].

SBA is a distance transform that is based on the distance maps (Euclidean distance) for each label in every atlas segmentation used. In the Euclidean distance map, voxels inside the labels have negative distances to the edge of the segment, and those outside have positive distances. The mean distance is

calculated for every voxel position ($\vec{x}$) to the current label (l) by summing the voxel values ($d_{k,l}$) for all the atlases and dividing it by the total number of atlases (K),

$$D_l(\vec{x}) = \frac{1}{K}\sum_{k=1}^{K} d_{k,l}(\vec{x}) \quad (1)$$

To get an output that consists of all the labels ($S(\vec{x})$), SBA is iteratively computed by minimizing the mean distance ($D_{\min}(\vec{x})$) over all labels like in Figure 4.

```
▷ Loop over all pixels to initialize data structures
for all x⃗ do
    ▷ Initialize output value to "reject"
    S(x⃗) ← L
    ▷ Initialize total distance map
    D_min(x⃗) ← ∞
end for
▷ Loop over all labels
for l = 0, ..., L−1 do
    ▷ Loop over all input images
    for all k = 1, ..., K do
        ▷ Compute signed distance map for current label
        d_{k,l} ← signed distance map of label l, image k
    end for
    ▷ Loop over all pixels
    for all x⃗ do
        ▷ Avg. signed distances for this pixel and label
        D ← (1/K) Σ_k d_{k,l}(x⃗)
        ▷ Is the new distance smaller than minimum?
        if D < D_min(x⃗) then
            ▷ Update combined label map
            S(x⃗) ← l
            ▷ Update minimum average distance
            D_min(x⃗) ← D
            ▷ Is the new distance equal to minimum?
        else if D = D_min(x⃗) then
            ▷ Mark pixel as ambiguous ("reject")
            S(x⃗) ← L
        end if
    end for
end for
```

Figure 4. Rohlfing's pseudocode representation of the SBA-algorithm [18].

The iterative algorithm in Figure 4 was implemented and applied segment by segment (i.e. 95 times) to combine the 30 atlas segmentations. The algorithm was written in Bash shell script using the Medical Image Registration Toolkit (MIRTK) [19] and the "seg_maths"-tool from the NiftySeg package [20].

SBA was used on the 7-Tesla image with 500 µm resolution. The segmentation done with SBA was visually compared with the VRF-segmentation of the same brain image. The Jaccard coefficient averaged over all labels between SBA and VRF was calculated to get a measure of the similarity between

the different types of fusing the segmentations of the target image acquired with MAPER. To get a better understanding of which one of SBA and VRF gives the most accurate segmentation, they were visually compared with two of the manually segmented brain images (#27 and #28).

Internal visual evaluation

The first step in evaluating the quality of the segmentations was to review the segmentations in the context of the MR image to assess the visual impression subjectively. Rview from MIRTK (called view in recent versions), a software that enables review of an MR image volume in three orthogonal planes, was used. Segmentations superimposed as label outlines differentiated by colour on the image made it possible to notice large differences from the brain anatomy. The disparity between the labels and brain anatomy was written down for all of the target images.

When comparing segmentations with various resolution, the 500 μm image with the segmentation superimposed as label outlines differentiated by colour was opened in rview. In another window of rview, the 250 μm images were opened one by one with their coloured labels from the 250 μm that were downsampled to 500 μm. Written notes were taken of the differences between the 500 μm segmentation and all of the segmentations of the 250 μm that were downsampled to 500 μm to see if they all differed the same from the original 500 μm segmentation. The 1 mm segmentation was only compared with the 500 μm segmentation in the same way as previously explained.

Visual evaluation by external experts

Four expert visual raters, chosen by the main supervisor's recommendation, were asked to rank a subset of the segmentations based on visual impression of quality on sample sections, and to explain how they were thinking. Table 1 lists the four images for which Pincram masks were shown, as well as the three images for which MAPER segmentations were shown. The questionnaire that were sent to the experts can be found at soundray.org/questions-experts.html.

The ranking points of the segmentations from all four experts were summed up and divided by the total number of experts to generate a mean score. The relative score was then calculated by dividing the mean score by the total amount of images of each slice.

Table 1. Lists of which images were shown for the experts to rank.

Pincram	**MAPER**
1 mm resolution	250 μm resolution, averaged
500 μm resolution	500 μm resolution
250 μm resolution, non-averaged	500 μm resolution, SBA
250 μm resolution, averaged	

Results

Ultrahigh-resolution image

The following images that have ultrahigh resolution are the original images with 250 μm cubic voxels, but with segmentations that have been acquired with the downsampled 500 μm version of the ultrahigh-resolution images. There was no visual difference between the original and downsampled images at the required zoom level. Comparison was made with all of the 8 images acquired at 5 different sessions. The one used to represent them below is the image acquired at the first session. The image that is called "the averaged image" in the following sections is the image that is an average of the 8 original images with 250 μm cubic voxels.

Pincram

When comparing the Pincram segmentations of the 9 downsampled images, the one with 500 μm resolution and the one with 1 mm resolution, significant differences were observed in the superior and inferior regions of the brain.

In the upper part of the brain, it was the downsampled, ultrahigh-resolution, non-averaged image that had the best-fitting brain mask to the parenchyma of the brain. The same was observed in all of the non-averaged images, and image ***c*** in Figure 4 shows one of them. In the averaged image (***d***), it could be seen that the brain mask deviated further from the cortex surface than in both the non-averaged images and the image with 500 μm resolution (***b***). In image ***a*** in Figure 5, it was the brain image with 1 mm resolution which excluded more of the sulci than the non-averaged image. It did seem to have a better fit to the parenchyma of the brain than both the 500 μm resolution image and the averaged one.

In the inferior brain regions, the segmentation acquired with the downsampled ultrahigh-resolution images, both the averaged (***d***) and non-averaged images (***c***), seemed to exclude a portion of cerebellar tissue compared to the segmentations from the 500 μm resolution (***b***) and 1 mm resolution images (***a***). The exclusion seemed to be greater for all 8 of the non-averaged images compared to the averaged. An axial slice of the lower part is represented in Figure 6.

Around the anterior part of the brain in the section shown in Figure 6, the structures were not as clear in the ultrahigh-resolution images as in the 500 μm resolution image. So the brain masks acquired with the downsampled non-averaged and averaged images also seemed to exclude some of that part of the brain. In the 1 mm resolution image, the delineation of the cerebellum included more voxels that did not contain brain tissue compared to the image with 500 μm resolution. In the frontal part of the brain, the 1 mm resolution image looked more accurate than the non-averaged and averaged images with 250 μm resolution.

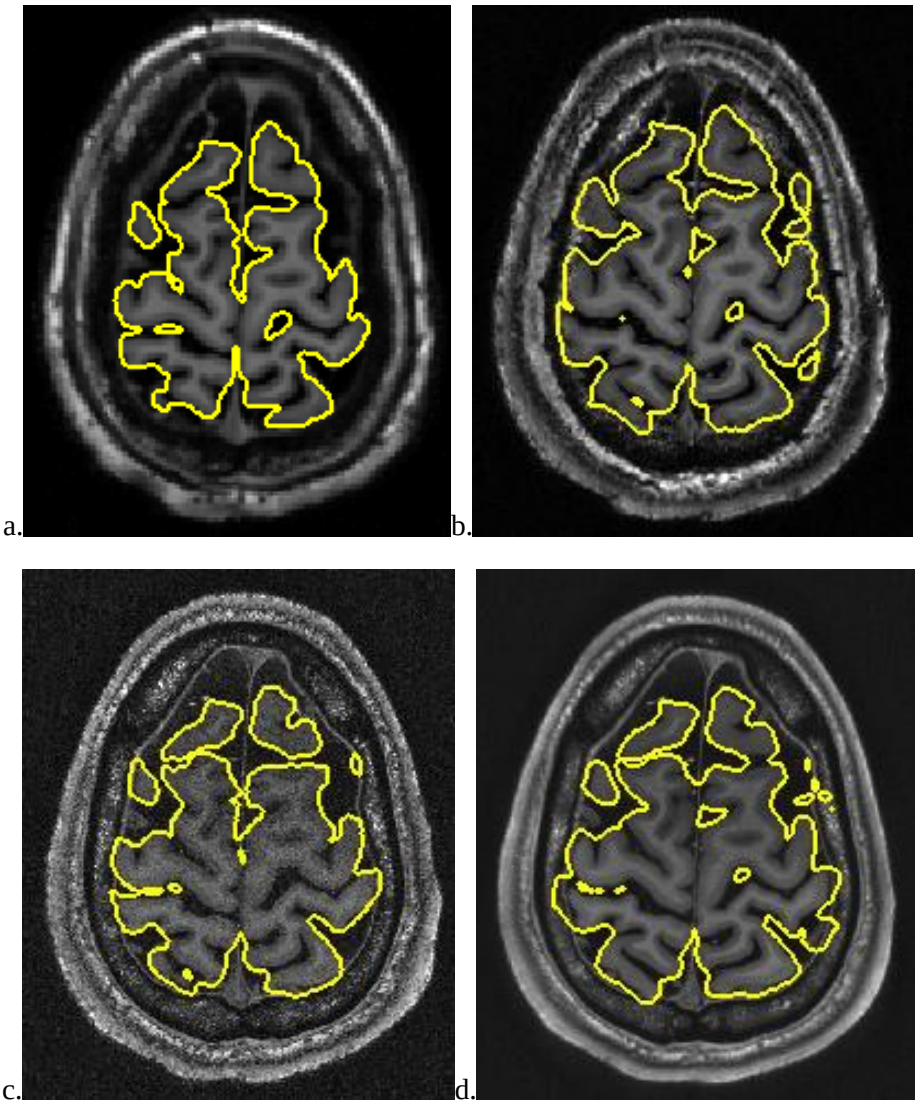


Figure 5. ***a.*** *Axial slice in the upper part of the brain image with 1 mm resolution and the brain mask acquired with Pincram.* ***b.*** *500 μm resolution image and the corresponding Pincram brain mask.* ***c.*** *One of the non-averaged images with 250 μm resolution and the Pincram brain mask acquired with the downsampled version of the same image.* ***d.*** *The averaged image with 250 μm resolution and the Pincram brain mask acquired with the downsampled version of the same image.*

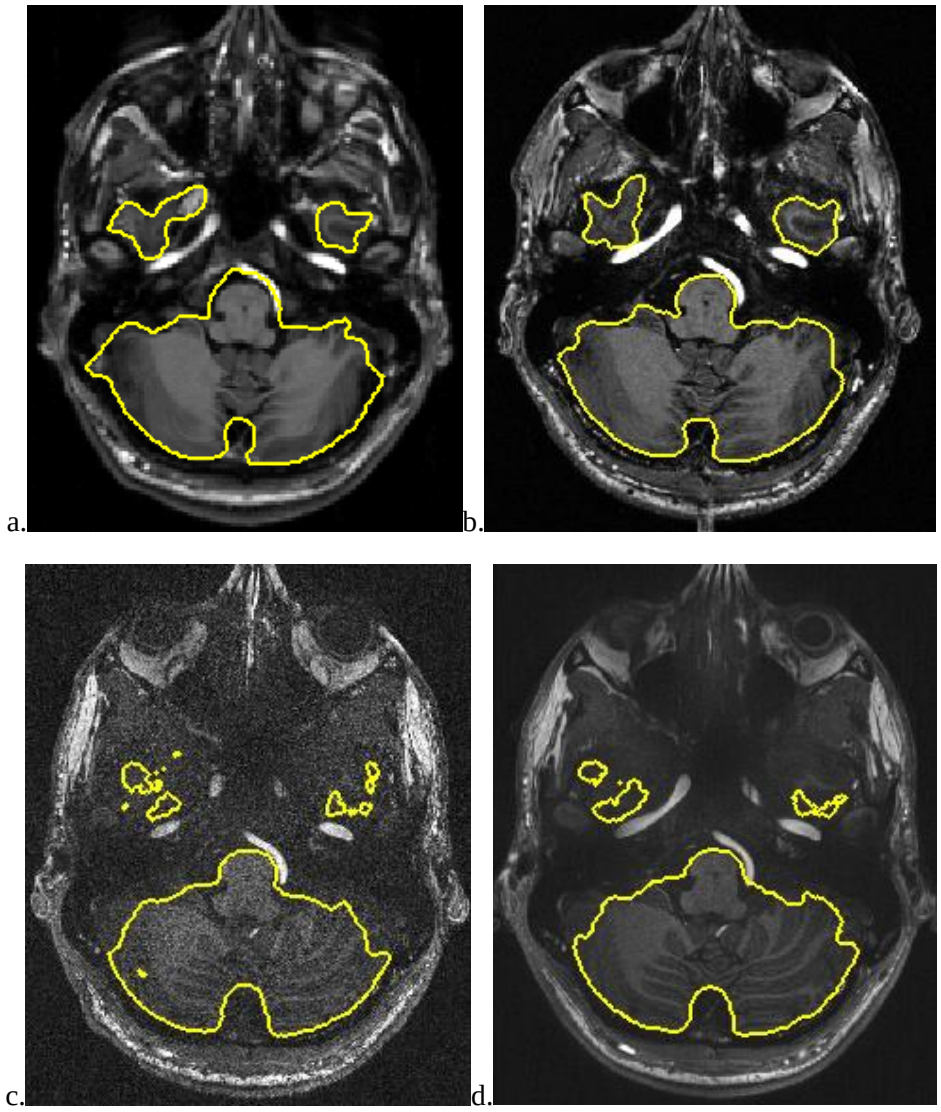


Figure 6. ***a.*** *Axial slice from the lower part of the brain image with 1 mm resolution and the corresponding brain mask acquired with Pincram.* ***b.*** *The 500 μm resolution image and the corresponding Pincram brain mask.* ***c.*** *One of the non-averaged images with 250 μm resolution and the Pincram brain mask acquired with the downsampled version of the same image.* ***d.*** *The averaged image with 250 μm resolution and the Pincram brain mask acquired with the downsampled version of the same image.*

In Table 2, the Jaccard coefficients for the overlap between the brain masks for all of the downsampled ultrahigh-resolution images are shown. 0.970 was the highest JC and it was between the two runs in Session 4. The lowest JC was 0.908 and represented the overlap between the first run in Session 2 and the first run in Session 5. The mean JC of all the calculated JC's in the table was 0.940.

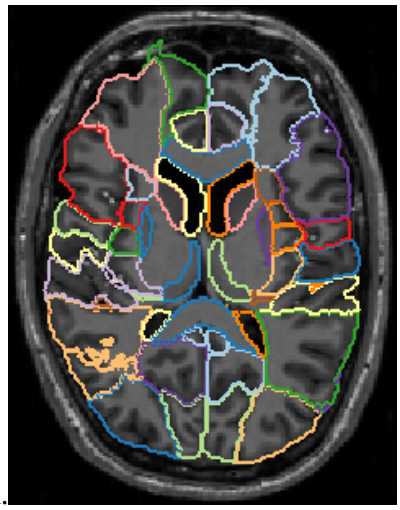
Table 2. The Jaccard coefficients for comparison of all of the MAPER-segmentations acquired with the downsampled ultrahigh-resolution images. s-01 to s-05 are session numbers as provided in the public dataset. The red numbers are the extreme values of the calculated Jaccard coefficients.

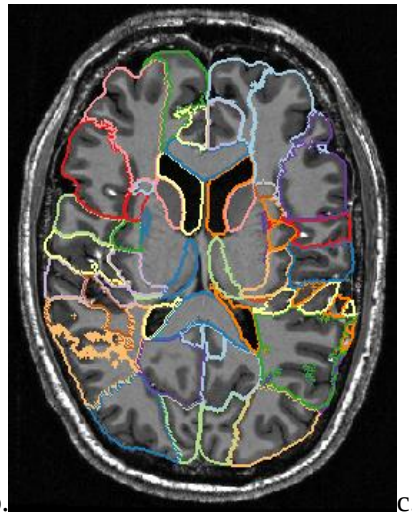
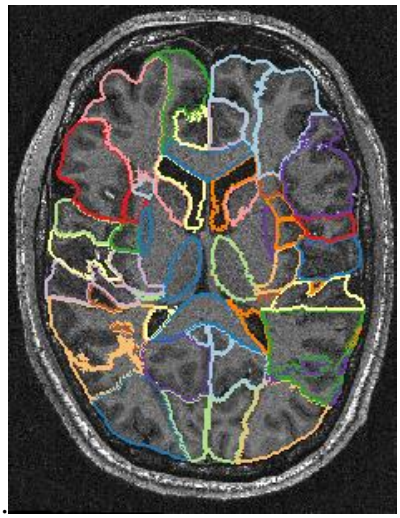
	s-01	s-02 run-01	s-02 run-02	s-03	s-04 run-01	s-04 run-02	s-05 run-01	s-05 run-02	Averaged Image
s-01	1	0.940	0.941	0.952	0.949	0.948	0.920	0.928	0.947
s-02 run-01	0.940	1	0.917	0.948	0.962	0.960	**0.908**	0.920	0.955
s-02 run-02	0.941	0.917	1	0.946	0.927	0.929	0.940	0.938	0.936
s-03	0.952	0.948	0.946	1	0.956	0.957	0.929	0.938	0.965
s-04 run-01	0.949	0.962	0.927	0.956	1	**0.970**	0.913	0.923	0.960
s-04 run-02	0.948	0.960	0.929	0.957	0.970	1	0.916	0.927	0.963
s-05 run-01	0.920	0.908	0.940	0.929	0.913	0.916	1	0.960	0.918
s-05 run-02	0.928	0.920	0.938	0.938	0.923	0.927	0.960	1	0.930
Averaged image	0.947	0.955	0.936	0.965	0.960	0.963	0.918	0.930	1

MAPER

When visually comparing the segmentation of the brain image with 500 μm resolution with the brain segmentation of the downsampled ultrahigh-resolution images, the most significant differences in the segmentations were found in the middle and inferior parts of the brain.

For the non-averaged images with ultrahigh resolution, the ventricles showed more irregular boundaries and a smaller volume than the image with 500 μm resolution (***b*** in Figure 7), the averaged image (***d*** in Figure 7) and the 1 mm resolution image (***a*** in Figure 7). Image ***c*** in Figure 7 shows one of the non-averaged images to show this effect. The same effect could be seen on the segmentation acquired with VRF of the 8 segmentations acquired with the downsampled non-averaged images (***e*** in Figure 7), but the boundaries for the ventricles were not as irregular as for the downsampled non-averaged image.

a. 

b. 
c. 

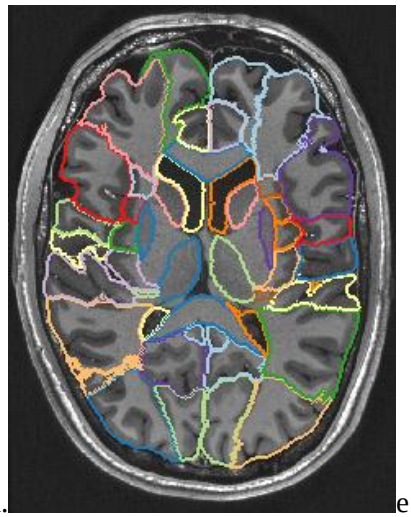
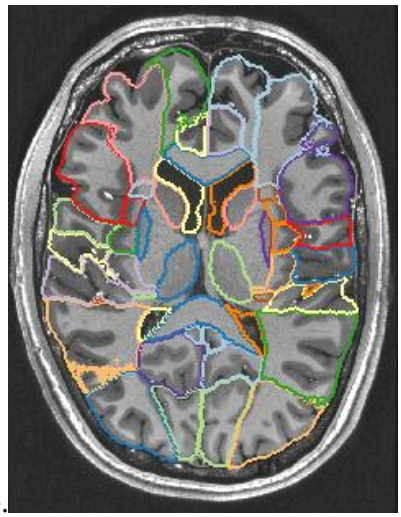
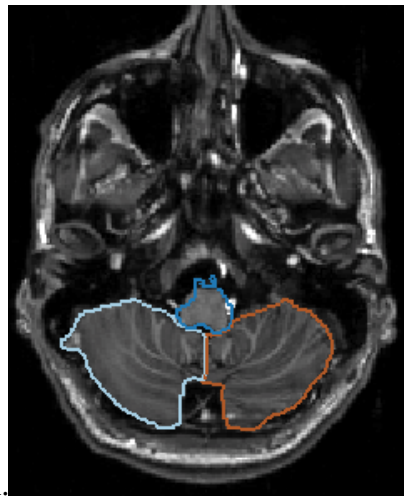
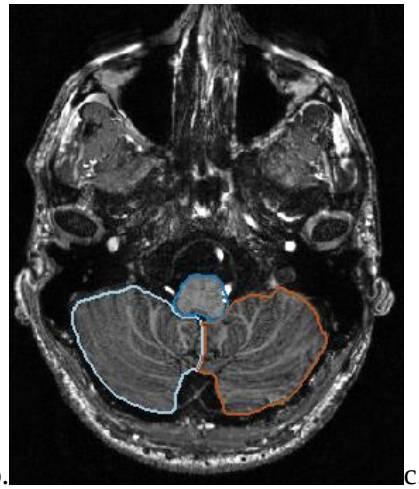
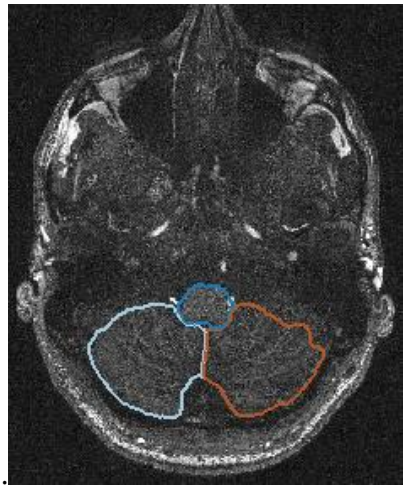
d. 
e. 

Figure 7. ***a.*** *Axial slice in the middle of the brain image with 1 mm resolution and the anatomical segmentation acquired with MAPER.* ***b.*** *The 500 μm resolution image with the corresponding MAPER-segmentation.* ***c.*** *One of the non-averaged images with 250 μm resolution and its corresponding MAPER-segmentation acquired with the downsampled version of the same image.* ***d.*** *The image that has been averaged from the 8 non-averaged images with 250 μm resolution and the MAPER-segmentation done with the downsampled version of the image.* ***e.*** *The averaged, ultrahigh-resolution image with the VRF of the segmentations of the 8 non-averaged images.*

The cerebellum and brainstem were delineated in the slice shown in Figure 8. For all 8 downsampled ultrahigh-resolution images (***c***), the size of the cerebellum was underestimated compared with the segmentation acquired with the 500 μm resolution image (***b***), the 1 mm resolution image (***a***) and the segmentation of the downsampled averaged image (***d***). The 8 non-averaged images did not underestimate the size of the cerebellum as much as the VRF of the 8 segmentations acquired with the downsampled, non-averaged images (***e***) did. So the best segmentation of this region of the brain was achieved with the 500 μm resolution image.

The mean calculated Jaccard coefficients for all 95 segments between all of the downsampled 250 μm resolution images can be seen in Table 3 in the light grey cells. The largest overlap was between run one and two in Session 4 with a mean JC of 0.836. The smallest overlap was between run one and two in Session 2 with a mean JC of 0.640. The mean of all the JCs in the table is 0.740. The darker grey cells in Table 2 contain the mean Jaccard coefficients for all 95 segments between the VRF of the 8 MAPER segmentations of the non-averaged images and the rest of the segmentations. For the VRF segmentation, the largest overlap was with the image acquired is Session 3 with a JC of 0.877 and the least overlap was with the first run of Session 5 with a JC of 0.769.

a. 

b. 
c. 

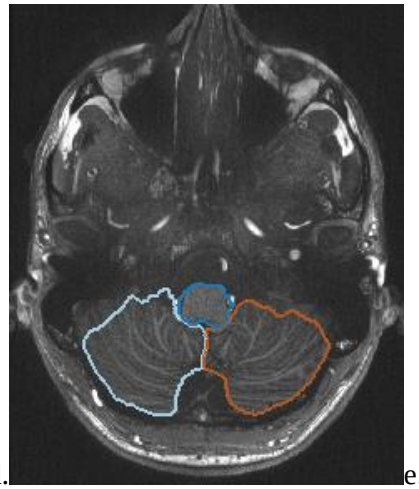
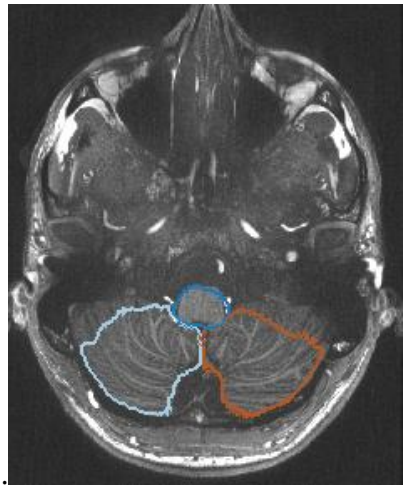
d. 
e. 

Figure 8. Axial sections in the region of the cerebellum. Light blue: right cerebellum; brown: left cerebellum; dark blue: brainstem. ***a.*** *Segmentation at 1 mm resolution.* ***b.*** *Segmentation at 500 μm resolution.* ***c.*** *One of the 250 μm images with the segmentation acquired with the downsampled version of the same image.* ***d.*** *The averaged 250 μm resolution image with the segmentation acquired with the downsampled version of the image.* ***e.*** *The averaged ultrahigh-resolution image with the VRF of the segmentations of the 8 non-averaged images.*

Table 3. The mean JC for all 95 segments between the 8 images from the non-averaged and the averaged image, all obtained on downsampled ultrahigh-resolution images. Extreme values of the Jaccard coefficients in the lighter grey cells are the red numbers. The darker grey cells contains the mean JC for all 95 segments between the VRF segmentation and the segmentations acquired from the rest of the downsampled images. The highest Jaccard coefficient over all is the yellow value.

	s-01	s-02 run-01	s-02 run-02	s-03	s-04 run-01	s-04 run-02	s-05 run-01	s-05 run-02	Averaged image	VRF of all 8 runs
s-01	1	0.720	0.724	0.772	0.759	0.758	0.693	0.707	0.747	0.824
s-02 run-01	0.720	1	**0.640**	0.764	0.788	0.778	0.646	0.679	0.749	0.789
s-02 run-02	0.724	0.640	1	0.734	0.687	0.702	0.780	0.752	0.708	0.776
s-03	0.772	0.764	0.734	1	0.791	0.799	0.729	0.752	0.788	**0.877**
s-04 run-01	0.759	0.788	0.687	0.791	1	**0.836**	0.671	0.711	0.810	0.840
s-04 run-02	0.758	0.778	0.702	0.799	0.836	1	0.690	0.731	0.817	0.857
s-05 run-01	0.693	0.646	0.780	0.729	0.671	0.690	1	0.792	0.690	0.769
s-05 run-02	0.707	0.679	0.752	0.752	0.711	0.731	0.792	1	0.740	0.803
Averaged image	0.747	0.749	0.708	0.788	0.810	0.817	0.690	0.740	1	0.818
VRF of all 8 runs	0.824	0.789	0.776	0.877	0.840	0.857	0.769	0.803	0.818	1

Shape-based averaging

Figure 9 illustrates the visual differences between VRF and SBA. The figure shows the three axial slices which contained the largest differences between VRF and SBA of the original 500 μm resolution image as judged visually. For all three slices, the left image was the one where the segments have been fused with VRF and the right with SBA. The overlap between the labels measured as an averaged JC over all labels was 0.89 between VRF and SBA.

In the first and most cranial slice, the biggest difference between the segmentations was the lines separating the superior frontal gyrus (darker green & light-blue) and the precentral gyrus (lighter green & red) on both sides of the brain. The label boundary voxels were much more scattered for the VRF than for the SBA segmentation. In general, the SBA-generated segments had smoother lines. In the second slice, there was one small and yellow segment that was surrounded by the green, superior frontal gyrus on the left side of the brain in the figure. The yellow region got much smaller with SBA than with VRF. In the third slice, the largest difference was in the posterior half of the brain where the lines between segments were much smoother for the SBA than the VRF segmentation.

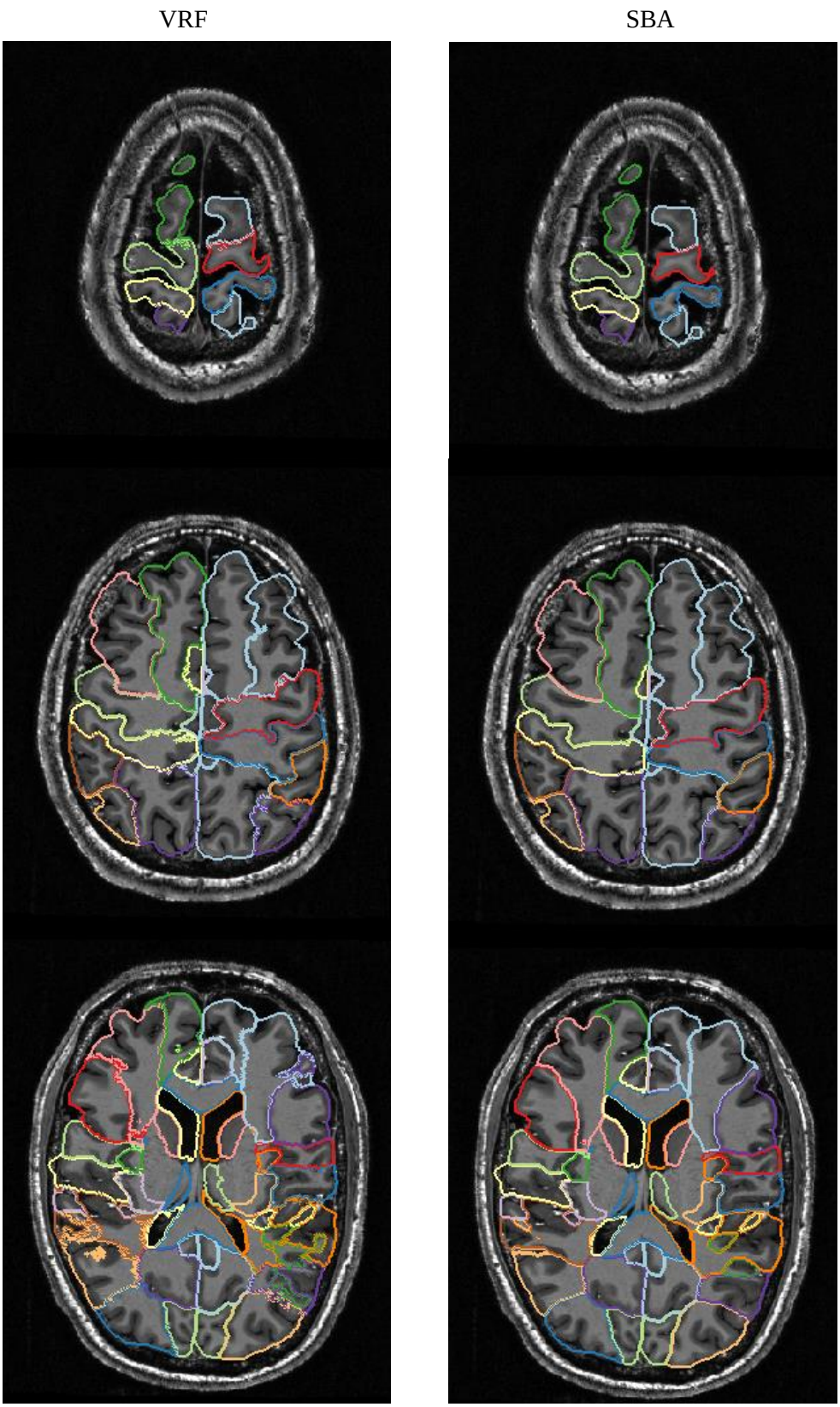


Figure 9. Three axial slices of the same brain image with 500 μm resolution. Anatomical segmentations generated with different methods are superimposed. The left segmentation is done with VRF and the right one with SBA.

Figure 10 shows the VRF segmentation on the left and SBA on the right. It is zoomed in to get a better view of how SBA diminishes the uncertainties in the boundaries between labels and smoothens the lines. In the uncertain regions, SBA tended to move the boundary between labels and cut off parts of one of the structures as in the red circles in Figure 10.

VRF SBA

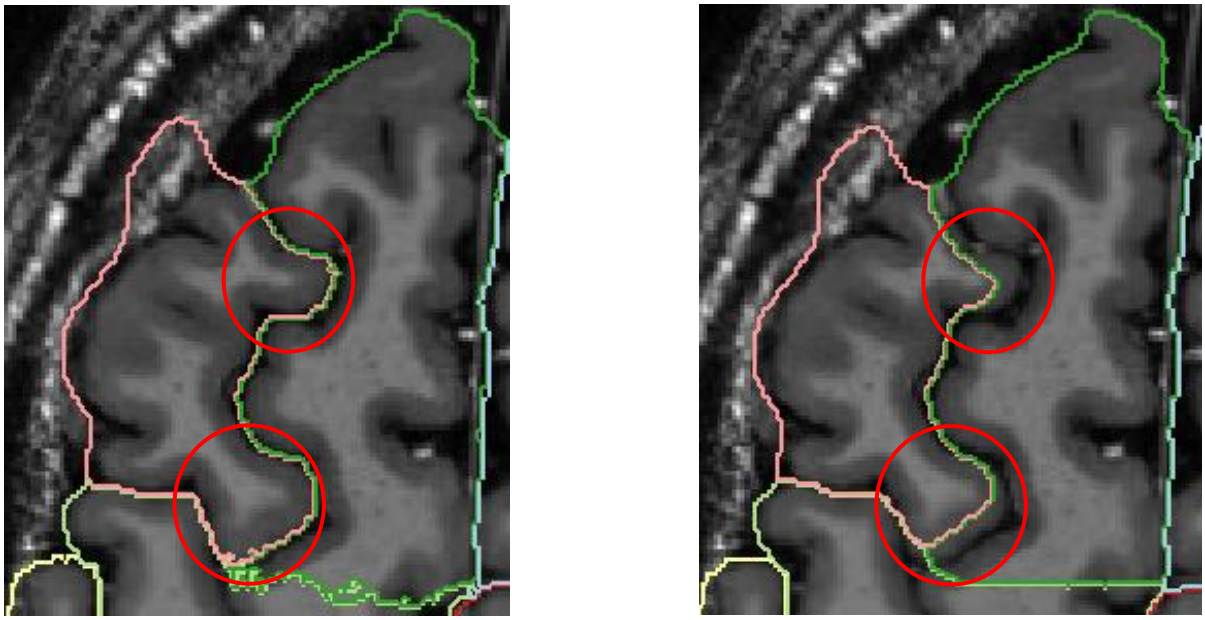


Figure 10. Zoomed-in images of the segmented brain with 500 μm resolution acquired at 7 Tesla. Segmentation done with VRF to the left and with SBA to the right. The red circles highlights the displacement of the lines between the structures for VRF compared to SBA.

Another phenomenon that was noticed in the bottom row of Figure 9 was that SBA sometimes splits a structure in two where the VRF segmentation was uncertain of which label the local structure belongs to. Figure 11 shows the zoomed in images of the splitting of a structure for SBA compared to the same region of the VRF segmentation. Several instances of this effect have been noticed all over the brain.

VRF SBA

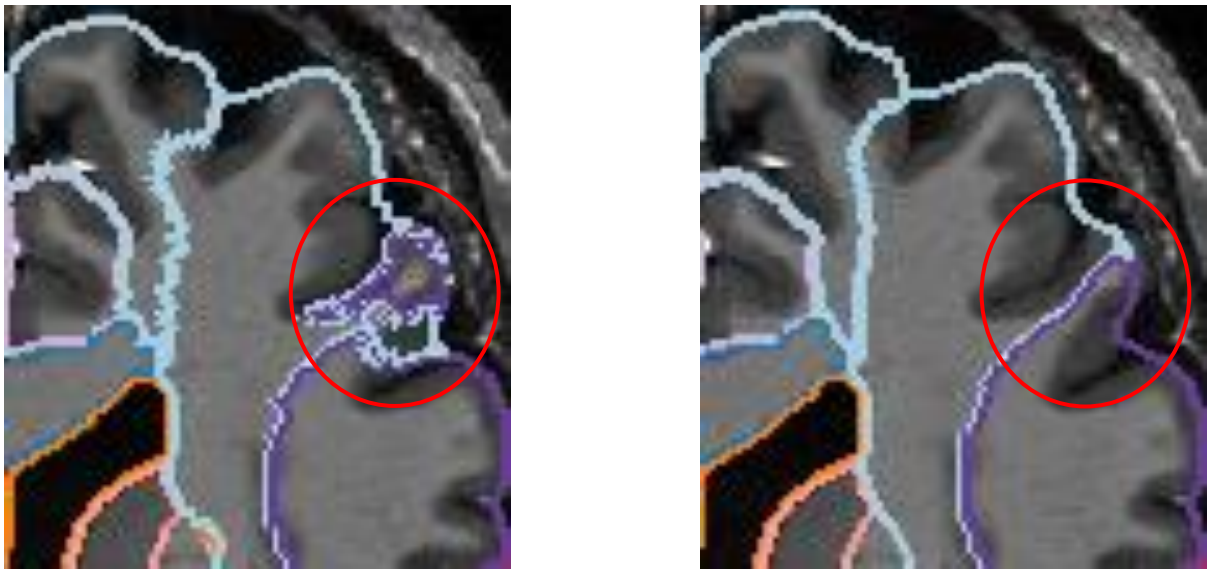


Figure 11. Zoomed in images of the segmented brain with 500 μm resolution acquired at 7 Tesla. Segmentation done with VRF to the left and with SBA to the right. The red circle highlights the splitting of the structure for SBA compared to VRF that has more irregularity between the segments.

Visual evaluation by external experts

The Pincram brain mask that received the lowest score (Table 4) of the ones that were shown in the questionnaire (Q1 in soundray.org/questions-experts.html) was the segmentation acquired with the 1 mm resolution image and therefore had the best visual impression of quality according to the experts. Their motivation for this decision was that the outer cortex delineation looked better than the other segmentations. The MAPER segmentation of the segmentations shown in the questionnaire (Q2 in soundray.org/questions-experts.html) according to the experts was the segmentation acquired with the downsampled averaged 250 μm resolution image because of the superior delineation of the thalamus compared to the other segmentations. The segmentation of the downsampled averaged image with ultrahigh resolution got the best score in Q2 and the worst in Q1.

Table 4. Individual ranking of each segmentation done by the experts and the mean score for each segmentation.

Pincram	Person1	Person2	Person3	Person4	Mean score (sum/#persons)	Relative score (mean/#images)
1 mm resolution	1	1	1	2	1.25	0.31
500 μm resolution	2.5	3	4	4	3.4	0.85
250 μm resolution, non-averaged	2.5	2	2	1	1.9	0.48
250 μm resolution, averaged	4	4	3	3	3.5	0.88
MAPER	Person1	Person2	Person3	Person4	Mean score (sum/#persons)	Relative score (mean/#images)
500 μm resolution	3	3	3	3	3	1
500 μm resolution, SBA	2	2	2	1	1.75	0.58
250 μm resolution, averaged	1	1	1	2	1.25	0.42

Discussion

MR scanning with 7 Tesla opens up new possibilities to study the anatomy of the human brain due to the high resolution achievable. There are some challenges using a 7 Tesla scanner for brain scans. The loss of signal in the inferior part of the brain due to the absence of a body coil led to lower SNR in the part where the fine structure of the cerebellum would need less noise to clearly be seen.

The present study is an investigation of automatic multi-atlas based anatomical segmentation methods applied to structural MR brain images acquired at 7 Tesla. It seems to be the first evaluation of its kind. The in-house developed algorithm that was used had previously only been applied to images acquired at 1.5 and 3 Tesla for clinical research purposes [21]. It was therefore unclear how well the algorithm and the atlases containing images acquired at 1.5 Tesla would work on images acquired at 7 Tesla. Would it even be possible, or would it require changes in the software or even new atlases?

Thanks to the extensive 7-Tesla data publicly shared by Lüsebrink et al., high quality 7-Tesla images with high resolution were available for this study. Their work shows that it is possible to obtain MR images with ultrahigh resolution without any major artefacts due to their advanced image processing, in particular their method for motion correction. Their comprehensive work with image processing made the present work easier because images were ready to be used immediately, rather than having to develop a preparation pipeline before the segmentation.

When applying SBA to the segmentation acquired with the 500 μm resolution image, the boundaries between the labels got much smoother and more pleasant to look at than with VRF.

The experts could tell if a specific structure was well segmented or poorly segmented without knowing which structure it was supposed to delineate which means that the algorithm is doing its job. The level of agreement between the experts was strong and they ranked the segmentations nearly identical.

Ultrahigh-resolution image

One of the segmentations acquired with one of the non-averaged images was shown in every Figure in the result. The reason for that was because they all had the same characteristics and underinclusions. The slices from the different acquisitions were approximately the same. Because of the different number of voxels in the 500 μm resolution image and the 250 μm resolution image, anatomical marks were used to find the same region of the images instead of the number of the current slice.

MAPER segmentations were strongly conditioned on the brain mask generated with Pincram, therefore the most outer lines were almost the same as the corresponding Pincram brain mask for all of the images in Figure 7 and 8.

The cerebellum turned out to be a problem region when Pincram and MAPER were applied to the downsampled ultrahigh-resolution images. It was distinctly underestimated, with underinclusions toward the anterior and inferior parts. On images with 1 mm and 500 μm resolution, this problem was also present, but on a smaller scale. Again, the problem appeared at the Pincram stage. MAPER relies on the corresponding brain mask, which places the cerebellar boundary inside the cerebellum (Figure 6), probably because the noise level in the non-averaged ultrahigh-resolution images increases substantially in the superior-inferior direction, and more so than in the 500 μm image.

With available software, it was not possible to calculate the JC between the segments that had been acquired on images with different resolution. That is why the JC was only calculated for the downsampled 250 μm, both non-averaged, averaged and the VRF of the 8 MAPER segmentations of the non-averaged images. The mean JC for all non-averaged images and the averaged image was higher for Pincram (0.940) than for MAPER (0.740). This is because of the size of the segments. Smaller segments and higher surface-to-volume ratio as in the MAPER segmentation generates a lower JC than

for the Pincram brain mask. Both for Pincram and MAPER, the highest JC was between the two runs in the fourth session which suggests that the time between the scans has an impact of the anatomy of the brain. There are, however, various variables that could have changed between scans and are hard to control. A case in point is that the lowest JC for MAPER was between the two runs in the second session. The difference between segmentations might have to do with the positioning and the movement of the subject, which have a greater impact on the images with high resolution. It is clear that the segmentations varied a lot because of the relatively low JC. Also, the range from the lowest to the highest JC was quite large for MAPER, with an interval of almost 0.2. For Pincram, the range interval was modest at 0.06.

The highest JC for the VRF of the MAPER segmentations of the 8 non-averaged images was 0.877 with the MAPER segmentation acquired with the image from Session 3. This does not say much of how accurate the segmentation of the VRF version was, but it shows that the image that looks the most like the VRF of the segmentations of the non-averaged images was the one from the third session. In Figure 7 and Figure 8, it can be seen that the VRF did not delineate the parenchyma as tightly as for the other images in the middle part of the brain. It also underestimated the size of the cerebellum more than the segmentation of the averaged image. That is why I choose not to show this segmentation to the experts.

Even though the segmentations were acquired with the downsampled images and not the original ultrahigh-resolution images, the segmentations would probably not have been better with the original ultrahigh-resolution images than the 500 μm resolution image. The reason for this speculation is that there was no visual difference between the downsampled and the original ultrahigh-resolution images. That is why the segmentations acquired with the downsampled images were assessed on the original ultrahigh-resolution images. What happens when doing so is that the segmentation only changes every other slice when scrolling through the volume. So, there would probably have been a difference in the segmentations of the original ultrahigh-resolution images compared to the downsampled, but the segmentations would probably not have been as good as for the 500 μm resolution image.

SBA

In the segmentation of the 500 μm resolution image acquired at 7 Tesla, irregular and scattered boundary voxels were noticed in several places. The manual segmentation protocols were formulated to imply that region labels should be continuous, there should be only one connected component per region. Visible irregularities in an automatic segmentation, therefore, suggest that something is wrong. The irregularities also distract the viewer, meaning that the labels stop being helpful for the viewer's anatomical orientation in the brain image.

Rohlfing et al. showed that shape-based averaging produces more regular and contiguous structures than VRF [18]. These results prompted me to implement SBA and to test whether the resulting labels would appear more consistent to the human eye and easier to distinguish from their neighbours. Indeed, applying SBA instead of VRF when segmenting the 500 μm resolution MR brain image acquired at 7 Tesla with MAPER resulted in smoother lines between the segments as shown in Figure 9, just as Rohlfing et al. suggested.

Between certain segments, the boundary was, however, clearly misplaced when using SBA (Figure 10). The reason for this misplacement may be that it was a strongly curved shape with large variations across the 30 atlases used for the MAPER segmentation. This means that the mean distance map of all 30 different segmentations of the pink area got smaller values than the green area in the region with grey matter in this case. Unlike in VRF, in SBA even the least similar atlas segmentation contributes to the final segmentation. Rohlfing et al. also claim that SBA has trouble dealing with high levels of intensity noise and requires the images to have the same grey level [18]. That might be a part of the reason why the fusion with SBA does not get as close to the fusion with VRF as wished for.

The fact that SBA divided the gyrus in Figure 11 is not so surprising. Figure 1 shows the corresponding section in one of the atlases. The appearance of the boundary dividing the structure is similar. This is

because the blue segment (left middle frontal gyrus) transitions into the purple segment (left inferior frontal gyrus) while scrolling from the most cranial slice and down. The transition became much smoother looking in the SBA segmentation than in both the VRF and the manual segmentation. This does not necessarily mean that it is the best volume estimation of the segments.

Visual evaluation by external experts

It was hard to draw any conclusions about which was the most accurate Pincram brain mask in the upper part of the brain (Figure 5) and the MAPER segmentation of the middle part of the brain (Figure 7). That is why some of them were to show to experts, to get a sense of which image and method are the best when segmenting the brain.

The segmentation of the downsampled averaged image with ultrahigh resolution was included in both Q1 and Q2 in the questionnaire. It received the best rating in Q2 and the worst in Q1, which means that the best image to use for segmentation of the brain with Pincram may be different from the best image to use with MAPER.

All four experts commented that it was challenging to rank the images. Their answers were nevertheless quite similar, which indicates that the assessment results are valid and that some image types were clearly better than others. Also, the fact that experts had trouble deciding on a ranking means that no segmentation had totally failed and that further evaluation of all segmentations is justified.

Another thing that two of the experts noticed was that the subject had unusually large sulci for a healthy person under the age of 40, implying that the subject has some degree of brain atrophy. This might affect the result of atlas-based brain segmentation if atrophy-related anatomical variation is not represented in the atlas. However, Pincram and MAPER have been designed to be robust for use in studying neurodegeneration and have been extensively used on, for example, images of elderly subjects with Alzheimer's disease [6, 22]. It is an indication that the accuracy of the segmentations was probably not affected by the atrophy noted.

The questionnaire that was shown to the experts only included axial slices of the brain which makes it harder to obtain a visual impression of the quality of the segmentations. The fact that it only showed a single brain section per question also made ranking the segmentations hard. The level of agreement between the experts is encouraging, but it is possible that a differently designed questionnaire would have yielded different results. For example, it would be desirable to show the whole brain image so that the segmentation of the whole brain could be evaluated. Doing a more comprehensive quality evaluation would be optimal, but would have exceeded the scope of this project.

Next step

Two things that neither SBA nor VRF take into account when fusing the segmentations that are acquired with MAPER using the atlases are the texture of the structures and the intensity in the image. A group of students at Chalmers University of Technology is currently working on incorporating texture attributes in the process of fusing the segmentation. Texture attributes are obtained by comparing the neighbouring voxels to each other.

To be able to run the ultrahigh-resolution images it would be beneficial to parallelize the image registration algorithms so that the process time and memory demands can better be adapted to the available hardware. The comparison between the segmentation of 7-Tesla images with the segmentation of 3-Tesla images exceeded the scope of this thesis but would be interesting to carry out due to the artefacts arising when using 7-Tesla scanners.

Conclusion

MAPER and Pincram work on brain images obtained with a 7-Tesla scanner even though the algorithms have been designed for and validated on 1.5 and 3 Tesla. The data size at the highest resolution exceed available computational resources, therefore images had to be downsampled. To be able to process the ultrahigh-resolution images without downsampling, the algorithms will need to be adapted to cope with the large information content generated.

SBA has been implemented on the MAPER-segmented atlases and applied to a 500 μm resolution image. The segment boundaries were smoother than with VRF and it got more pleasant to look at. Some boundaries do get misplaced, so the volume estimation of the structures might not be better than with VRF.

Acknowledgments

I would like to thank my supervisors Emilia and Rolf. Emilia for the helpful revisions of my report and Rolf for helping me through this project and teaching me all I know about atlas-based segmentation.

I also would like to thank Professor Isabella Björkman-Burtscher who sat down with me and explained what she thought of the segmentations and taught me more about the limitations with 7-Tesla scanners.

Last but not least Professor Hans-Jürgen Huppertz, Professor Alexander Hammers and Professor Stephan Maier for helping me evaluate the segmentations.

Reference list

1. Gonzalez, R.C. and R.E. Woods, *Digital Image Processing*. Third Edition ed. 2008, Pearson Education, Inc. 909.
2. Grove, E. *1st clinical 7-Tesla MRI scanner in North America will enable Mayo Clinic to achieve higher-resolution imaging for patients*. 2017; Available from: https://newsnetwork.mayoclinic.org/discussion/1st-clinical-7-tesla-mri-scanner-in-north-america-will-enable-mayo-clinic-to-achieve-higher-resolution-imaging-for-patients/.
3. Heckemann, R.A., et al., *Automatic volumetry on MR brain images can support diagnostic decision making.* BMC Medical Imaging, 2008. **8**(1).
4. *The progression of Alzeimer's disease*. 25 May 2017; Available from: https://www.brightfocus.org/alzheimers/infographic/progression-alzheimers-disease.
5. Heckemann, R.A., et al., *Automatic anatomical brain MRI segmentation combining label propagation and decision fusion.* NeuroImage, 2006. **33**(1): p. 115-126.
6. Heckemann, R.A., et al., *Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation.* NeuroImage, 2010. **51**(1): p. 221-227.
7. Heckemann, R.A., et al., *Brain Extraction Using Label Propagation and Group Agreement: Pincram.* Plos One, 2015. **10**(7).
8. Faillenot, I., et al., *Macroanatomy and 3D probabilistic atlas of the human insula.* Neuroimage, 2017. **150**: p. 88-98.
9. Gousias, I.S., et al., *Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest.* Neuroimage, 2008. **40**(2): p. 672-684.
10. Hammers, A., et al., *Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe.* Hum Brain Mapp, 2003. **19**(4): p. 224-47.
11. *Brain Development*. Available from: http://brain-development.org.
12. Lüsebrink, F., et al., *T1-weighted in vivo human whole brain MRI dataset with an ultrahigh isotropic resolution of 250 μm.* Scientific Data, 2017. **4**.
13. collaboratory, N.t.r. *Automatic Registration Toolbox*. 2008; Available from: https://www.nitrc.org/projects/art/.
14. Johnson, H.J., N.J. Tustison, and B.B. Avants. *Advanced normalization tools*. Available from: http://stnava.github.io/ANTs/.
15. Grinstead, J.W., W. Rooney, and G. Laub, *The Origins of Bright Blood MPRAGE at 7 Tesla and a Simultaneous Method for T1 Imaging and Non-Contrast MRA.* International Society for Magnetic Resonance in Medicine, 2010. **18**: p. 1429.
16. Rueckert, D., et al., *Nonrigid registration using free-form deformations: application to breast MR images.* Ttansactions on medical imaging 1999. **18**: p. 712-721.
17. Kittler, J., et al., *On combining classifiers.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(3): p. 226-239.
18. Rohlfing, T. and C.R. Maurer, *Shape-Based Averaging.* IEEE Transactions on Image Processing, 2007. **16**(1): p. 153-161.

19. A., S. *Medical Image Registration ToolKit (MIRTK)*. 2018; Available from: https://mirtk.github.io.
20. [cited 2018 3. December]; Available from: https://github.com/KCL-BMEIS/NiftySeg.
21. Heckemann, R.A. *maper*. [cited 2019 3. January]; Available from: https://soundray.org/maper/.
22. Heckemann, R.A., et al., *Automatic morphometry in Alzheimer's disease and mild cognitive impairment.* Neuroimage, 2011. **56**(4): p. 2024-37.