The final publication is available at

http://dx.doi.org/10.1016/j.neuroimage.2016.11.003

Additional Information

# CERES: A new cerebellum lobule segmentation method

Jose E. Romero[1], Pierrick Coupé[2,3], Rémi Giraud[2,3], Vinh-Thong Ta[2,4], Vladimir Fonov[5],

Min Tae M. Park[6,7], M. Mallar Chakravarty[7,8], Aristotle N. Voineskos[9,10], Jose V. Manjón[1]

[1] Instituto Universitario de Tecnologías de la Información y Comunicaciones (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, España.
[2] Univ. Bordeaux, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France.
[3] CNRS, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France.
[4] Bordeaux INP, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France.
[5] BIC, Montreal Neurological Institute, Montreal, Canada.
[6] Schulich School of Medicine and Dentistry, Western University, London, Canada.
[7] Cerebral Imaging Centre, Douglas Mental Health University Institute, Montreal, Canada.
[8] Department of Psychiatry and Biological and Biomedical Engineering, McGill University, Montreal, Canada.
[9] Campbell Family Mental Health Institute, Centre for Addiction and Mental Health, Toronto, Canada.
[10] Centre for Addiction and Mental Health, Toronto, Canada.

## Abstract

The human cerebellum is involved in language, motor tasks and cognitive processes such as attention or emotional processing. Therefore, an automatic and accurate segmentation method is highly desirable to measure and understand the cerebellum role in normal and pathological brain development. In this work, we propose a patch-based multi-atlas segmentation tool called CERES (CEREbellum Segmentation) that is able to automatically parcellate the cerebellum lobules. The proposed method works with standard resolution magnetic resonance T1-weighted images and uses the Optimized PatchMatch algorithm to speed up the patch matching process. The proposed method was compared with related recent state-of-the-art methods showing competitive results in both accuracy (average DICE of 0.7729) and execution time (around 5 minutes).

**Keywords:** MRI, cerebellum lobule segmentation, non-local multi-atlas patch-based label fusion, optimized patchmatch.

# 1. Introduction

The human cerebellum is a neuroanatomical structure within the human brain located below the cerebrum and connected to the brainstem through the cerebellar peduncles. Although it represents a small percentage of the total intracranial volume, about 10%, it plays a key role in motor coordination and learning (Manto et al., 2013). The cerebellum acts as a modulator for the motor cortex signals to perform fine and precise movements through the integration of the inputs signals from the somatosensory system and cerebellar cortex. It has been shown that cerebellar damage leads to important motor dysfunctions such as limb ataxia, balance alterations and other deficits in motor coordination (Kase et al., 1993; Compston and Coles, 2008; Davie et al., 1995; Klockgether, 2008). Recent neuroimaging studies indicate that cerebellum is also involved in many cognitive functions (Stoodley, 2012) such as attention, emotion or behavior (Timmann et al., 2010).

Despite the cerebellum being involved in many brain functions, it has been widely understudied. Lately, several publications have focused on the study of the cerebellum and its implication in several neuropsychiatric pathologies. For example, patients with schizophrenia showed a reduced vermis (Okugawa et al., 2002) and subjects with Alzheimer's disease presented smaller posterior lobes (Thomann et al., 2008). Nevertheless, there exist publications with inconclusive results regarding the role of the cerebellum in different neuropsychiatric disorders (Nenadic, Sauer and Gaser, 2010; James et al., 2004). This could be explained by the fact that accurately segmenting the cerebellum is quite challenging due to its complex anatomy.

Cerebellar anatomy consists of a white matter tree structure located behind the pons. It is divided into two hemispheres (left and right), with each white matter branch surrounded by a layer of grey matter that creates folds called foliations. These grey matter folds are denominated cerebellum lobules. The size, position and number of (visible) lobules is highly variable between subjects which makes the segmentation process challenging.

Manual segmentation (delineation) by an expert requires a huge effort since it is a time consuming and tedious process. Delineation by non-expert raters may be faster but still impractical for studies applied to a large cohort of subjects. Further, both alternatives require dealing with inter and intra-rater variability. A few semi-automatic methods have been developed offering large structure parcellation as the brainstem, white matter and grey matter (Weier et al., 2012). These methods intend to reduce the time required by a human rater while keeping the reliability of manual delineation. However they are still time consuming.

An alternative to the manual segmentation is using multi-atlas based segmentation (Aljabar et al., 2009) which consists of transferring and combining multiple manual segmentations to the case to be segmented. One of the first automated algorithms specifically developed for cerebellum segmentation is SUIT (Diedrichsen, 2006). In this method, a set of manually labeled images was used to build a probabilistic atlas that was then warped into the target image space. However, SUIT is not able to capture all the anatomical variability due to the use of a single atlas approach.

In addition, the method developed by Diedrichsen is semi-automatic since it includes a cerebellum extraction step that is suggested to be manually corrected for optimal results.

Fully automatic methods appeared in the last years. Bogovic et al. (2013) published a method called ACCLAIM using deformable models and random forest boundary classification. Recently Yang et al. (Yang et al., 2015) proposed a segmentation method combining multi-atlas label fusion and tissue/boundary classification in a graph-cut segmentation framework to address the high variability of the cerebellum anatomy achieving good results at the expense of a high computational burden (execution time of several hours per case). Also recently, a new approach based on multi-atlas segmentation called MAGeT Brain (Chakravarty et al., 2013; Park et al., 2014) has been proposed for cerebellum lobule segmentation. This new approach aims at solving the problem of having very few exemplar cases (manual segmentations) of a given segmentation protocol. It uses the estimation of a large set of non-linear deformations to map the example cases to a set of unlabeled templates increasing the number of candidate segmentation to apply a final majority vote based label fusion. MAGeT Brain demonstrated better results than single atlas and classical multi-atlas label fusion using majority voting methods at the expense of a high computational cost in terms of both memory and execution time. Finally, Weier et al. published another multi-atlas segmentation method named RASCAL (Weier et al., 2014) based on a multi-atlas patch-based label fusion. It utilizes a library of manually segmented cases that is non-linearly registered into the target image space to perform the labeling process.

In this paper, we introduce a new pipeline for cerebellum lobule segmentation that is based on an adaptation of a recently proposed segmentation method called Optimized PatchMatch Label fusion (OPAL) (Ta et al., 2014; Giraud et al., 2015). This method consists of a multi-atlas patch-based (Rousseau et al., 2011) segmentation with a non-local label fusion technique that produces fast and accurate segmentations using a library of manually segmented cases. The original method developed for hippocampus segmentation (Coupe et al., 2011) has been adapted to segment the cerebellum anatomy using a library of non-linearly registered cases instead of using only an affine registration as in the original method. A post-processing step has been added to enforce regularity of the different lobule labels. Also, given the small size of the template library (only 5 cases) we extended the number of templates of the library by adding automatically segmented cases.

## 2. Materials and methods

In the following sections, the data used in this paper and the proposed method details are described. We have called the proposed method CERES (for CEREbellum Segmentation).

### 2.1. Image data

In this work, two datasets have been used. A High Resolution (HR) set of 5 MR T1-weighted images has been used to construct the library of manually labeled cases. A second dataset of 20 standard resolution images has been used to evaluate the proposed method and to compare it with other related approaches. The HR images are publicly available at the CoBrALab website (http://cobralab.ca/atlases/Cerebellum.html). Both the HR images used as input and the manually labeled validation dataset are the same as those used in Park et al. (2014).

### HR dataset

These cerebellar atlases were created based on HR MR images from 5 healthy volunteers (2 male, 3 female, aged 29–57). We used T1-weighted images that were acquired on a GE Discovery MR 750 3 T system (General Electric, Waukesha, WI) using an 8-channel head coil. High-resolution T1-weighted images were acquired using the 3D inversion-prepared fast spoiled gradient-recalled echo acquisition, FSPGR-BRAVO, in a scan time of ~ 20 min with the following parameters: TE/TR = 4.3 ms/9.2 ms, TI = 650 ms, α = 8°, 2NEX, FOV = 22 cm, slice thickness = 0.6 mm, 384 × 384 in-plane steps for an approximate isotropic resolution of 0.6 mm dimension voxels. Reconstruction filters, ZIPX2 and ZIP512, were also used resulting in isotropic 0.3 mm$^3$ voxels with a final matrix size of 489x734x503 voxels (Park et al., 2014). These HR MR images were manually segmented by two expert raters that allowed for both inter- and intra-rater comparisons of segmentations to validate the consistency of the manual segmentation protocol (for details of the segmentation protocol please see (Park et al., 2014)). The cerebellum was manually delineated into 26 structures: White matter and Lobules I-II, III, IV, V, VI, Crus I, Crus II, VIIb, VIIIa, VIIIb, IX, and X considering left and right hemispheres as shown in Figure 1. One of the authors of this manuscript, Min Tae M. Park, was one of the manual raters of these data.
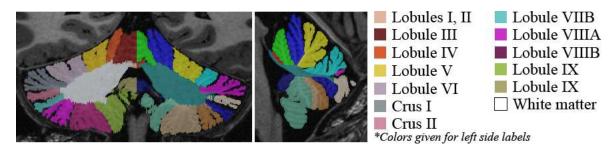


| | | | |
|---|---|---|---|
| Lobules I, II | | Lobule VIIB | |
| Lobule III | | Lobule VIIIA | |
| Lobule IV | | Lobule VIIIB | |
| Lobule V | | Lobule IX | |
| Lobule VI | | Lobule IX | |
| Crus I | | White matter | |
| Crus II | | | |

*Colors given for left side labels*

*Figure 1: Example of one HR case and its corresponding manual segmentation.*

**Standard resolution dataset**

MR images from 10 healthy control and 10 patients with schizophrenia (15 male, 5 female, average 30.1 years age, range 25–35 years) were used to create a validation set. Imaging data were acquired using an 8-channel head coil on a 1.5 T GE Echospeed system (General Electric Medical Systems, Milwaukee, WI), which permits maximum gradient amplitudes of 40 mT/m. Axial inversion recovery prepared spoiled gradient recall images were acquired: echo time (TE) = 5.3 ms, repetition time (TR) = 12.3 ms, time to inversion (TI) = 300 ms, flip angle = 20°, number of excitations (NEX) = 1 (124 contiguous images, 1.5 mm thickness, 0.78 x 0.78 mm in plane resolution). These images were linearly registered to the MNI space using MRITotal software (Collins et. al, 1994) to perform the manual labeling thus resulting in a 1 mm$^3$ image resolution and a matrix size of 181x217x181 voxels. The whole cerebellum is manually segmented on these 20 cases dataset and the left cerebellum was segmented on 10 of the cases while in the other 10 the right cerebellum was segmented (24 lobules: I-II, III, IV, V, VI, Crus I, Crus II, VIIb, VIIIa, VIIIb, IX, and X). The definition of the lobules is the same as the atlases and it is shown in Figure 1. One of the authors of this manuscript, Min Tae M. Park, was one of the manual raters of these data.

## 2.2. Preprocessing

All the images are preprocessed to enhance their quality and to transform them to a common intensity and geometric space. The steps included in this preprocessing are the following:

*Denoising*

The Spatially Adaptive Non-local Means Filter is applied to reduce the noise in the images. This filter was chosen because it is able to automatically deal with both stationary and spatially varying noise levels (Manjón et al., 2010).

*Inhomogeneity correction in native space*

The N4 bias field correction (Tustison et al., 2010) is applied to correct intensity inhomogeneities across the images. This step is applied to improve the registration process but not used as a final bias correction method as will be described later.

*MNI affine registration*

All the images have to be registered to a common coordinate space so that the anatomy within the case to be segmented and the library cases is consistent. To this end, the images are linearly registered (affine transform) to the Montreal Neurological Institute (MNI) space using the MNI152 template. This is done using the Advanced Normalization Tools (ANTs) (Avants, Tustison and Song, 2009).

*Inhomogeneity correction in MNI space*

N4 method is used to remove the inhomogeneity of the images in the MNI space (Tustison et al., 2010). We use the MNI125 intracranial mask to better correct the brain region. We found

experimentally that the application of N4 method in MNI space has a superior performance compared to its application on native space.

*Cerebellum area cropping*

A cropping step is applied to reduce the computational burden of the method limiting the processing to the cerebellum area. A cropping 3D window was calculated from the limits of the average of all the atlas labels of the library in the MNI space (including a margin of 5 voxels in all 3 dimensions).

*Low dimensional non-linear registration estimation*

To achieve a better cerebellum anatomic matching between the target image and the library of manually labeled templates, the cropped area of both (target and templates) is non-linearly registered to the cropped MNI152 atlas (the non-linear transformation is estimated but not applied). Although the segmentation process could be performed at this non-linear MNI152 space, the inverse transformation needed to map back the segmentation result on the original linear MNI space would reduce segmentation quality due to the introduction of interpolation artifacts.

Applying successive transformations implies performing successive interpolations. Instead, we segment the target case in the linear MNI space avoiding the use of a non-linear inverse transformation on the result. To this end, a customized subject-specific library is created by concatenating the forward non-linear transformation to the MNI152 atlas of each library case and the inverse non-linear transformation of the subject to be segmented. By doing this we are applying the transformation once, so only one interpolation is needed. Although the custom library has interpolation errors these are minimized during the label fusion step.

The registration step has a huge impact on segmentation accuracy. We tested 3 different resolution levels (low=4 mm$^3$, medium=2 mm$^3$ and high=1 mm$^3$). In general, a high dimensional transformation (1 mm$^3$) is expected to better match the reference anatomy. However, this is highly dependent on the data topology and we found that the low dimensional registration (4 mm$^3$) provided the best segmentation results (being also the most efficient in terms of computational load). The non-linear registration was performed using ANTs tool (Avants, Tustison and Song, 2009) with cross correlation similarity measure.

*Intensity normalization*

As previously commented, it is necessary to normalize the images in order to obtain the same intensity values across subjects for cerebellar tissues (white matter and grey matter). This is required to perform the patch-based segmentation that finds matches between the target image and the library of templates using an intensity-based similarity metric. To this purpose, a local intensity normalization based on regions of interest (ROIs) is applied in a similar manner to the method proposed by Asman and Landman (2012) in the multi-atlas segmentation challenge of MICCAI 2012.

We use the average MNI152 atlas as the reference image that is mapped to the target space by applying the previously computed non-linear inverse transformation. Then the majority voting segmentation obtained from the library is used to define several ROIs (one for each label). Next, the median intensity for each ROI is calculated for both target and atlas and a second order polynomial function is fitted to map the corresponding median intensities. Finally, the intensities of the target case are normalized to the atlas intensities by evaluating the fitted function at every target voxel intensity value. Note that we use median intensity values instead of mean values because it was found to be more robust to the presence of outliers.

## 2.3. Library construction

To construct the library of templates, the HR dataset mentioned in Section 2.1 was preprocessed as described in Section 2.2. Label maps delineated on native space were linearly mapped to the MNI space. Then, images and labels were left-right flipped to double the library size (from 5 to 10 labeled cases). The fact that the cerebellum is nearly symmetric allows the segmentation process to take benefit from this step as long as we are increasing the number of anatomical references (patterns) locally. Images and labels were then cropped to the cerebellum region. Afterwards, the non-linear transformation of these cropped images to the cropped MNI atlas was estimated. These transformations are used to non-linearly register the entire library to the space of the case to be segmented in a similar way as done in (Weier et al., 2014). Finally, the images were intensity normalized. Note that the resulting library at MNI space has a 1 mm$^3$ resolution (original HR data had a 0.3 mm$^3$ voxel dimensions).

## 2.4. Segmentation pipeline description

### Preprocessing

Any new unlabeled case is preprocessed as previously described by applying the denoising, the linear registration to the MNI space, the inhomogeneity correction, the cropping, the non-linear registration and the local intensity normalization. Finally, the estimated non-linear transformation is used to create the subject-specific library. After this process, all the library templates are non-linearly registered to the target case requiring only one non-linear transformation estimation instead of $N$ non-linear transformations (where $N$ is the number of subjects in the library).

### Labeling

*Non-local patch-based label fusion*

Non-local patch-based label fusion is a multi-atlas segmentation technique that combines multiple reference atlas segmentations to produce a new segmentation that fits the anatomy of a new unlabeled case (Coupé et al., 2011). In this strategy, the final label for each image voxel is estimated by performing a weighted label fusion of a set of sample patches within a specific search area centered on the target voxel for all N cases in the library. The estimation is computed as follows:

$$v(x_i) = \frac{\sum_{s=1}^{N} \sum_{j \in V_i} w(x_i, x_{s,j}) y_{s,j}}{\sum_{s=1}^{N} \sum_{j \in V_i} w(x_i, x_{s,j})}$$ (1)

where $V_i$ represents the search area, $N$ is the number of subjects in the template library, $y_{s,j}$ is a possible label from the voxel $x_{s,j}$ at the position $j$ in the subject $s$ and $w(x_i, x_{s,j})$ is the patch similarity defined as:

$$w(x_i, x_{s,j}) = exp^{\frac{-\left\|P(x_i) - P(x_{s,j})\right\|_2^2}{h^2}}$$ (2)

where $P(x_i)$ is the patch centered at $x_i$, $P(x_{s,j})$ the patch centered at $x_j$ in the templates and $||.||_2$ is the normalized L2 norm (normalized by the number of elements) calculated from the distance between each pair of voxels from both patches $P(x_i)$ and $P(x_{s,j})$. $h$ is a normalization parameter that is estimated from the minimum of all patch distances within the search area.

Although the non-local label fusion is a powerful segmentation technique, one of its drawbacks is its high computational cost (mainly due to the patch similarity estimation) that limits its use on large images and big libraries. Recently, a fast patch matching techniques have been proposed that significantly reduces the computational burden of non-local methods (Ta et al., 2014; Giraud et al., 2015).

*The PatchMatch algorithm*

The PatchMatch (PM) algorithm (Barnes et al., 2009) is an efficient strategy to find patch correspondences between two images (originally 2D). It is based on the hypothesis that given a patch $p(i,j) \in$ an image $A$ that matches a patch $p(i',j') \in$ an image $B$, then the adjacent patches of $p(i,j)$ will probably match the adjacent patches of $p(i',j')$.

This algorithm is implemented in three steps: initialization, propagation and random search. First, in the initialization, a random patch from $B$ is assigned to each patch in $A$. Second, adjacent patches are tested to improve the current match by following the concept of adjacency. Third, with the aim of avoiding local minima, random samples are selected within a neighborhood and compared to the current match. The second and third steps are repeated iteratively to improve the correspondences. This method quickly converges and gives an approximate nearest neighbor field (ANN) that is a set of correspondences between images $A$ and $B$.

*Optimized PatchMatch (OPAL)*

A 3D extension of PM to deal with MR images was recently proposed by (Ta et al, 2014). This method, called OPAL, applies PM to 3D MR images and finds patch correspondences between a given image $S$ and a library of $L$ templates. One of the benefits of the PM algorithm is that its complexity only depends on the size of the image $A$ and not on the size of the library $L$. This

advantage allows computation of patch correspondences between an image *S* and a large library *L* = *{T₁,...,Tₙ}* without using any template preselection at an almost constant computational cost (discarding library loading time). Compared to PM method, OPAL has a constrained random initialization since typically both target and library are located in the same geometric space (MNI space) and therefore good matches for the target patches are expected to be close in the library space. For further details, see the original OPAL paper (Ta et al., 2014) and its recent extension paper (Giraud et al., 2015).

*Multi-scale OPAL label fusion*

In patch-based segmentation, the patch size is a key parameter that is strongly related to the structure to be segmented and the image resolution. This can be empirically optimized for a specific dataset. It can be seen in the literature that multi-scale approaches improve segmentation results (Manjón et al., 2014). In OPAL (Giraud et al., 2015) independent and simultaneous multi-scale and multi-feature ANN fields were computed. In this paper, a multi-scale approach was also used but only voxel intensities are considered as a feature. Finally, several different ANNs are computed for different scales resulting in different label probability maps and have to be combined. In OPAL, the final estimation is obtained by simply averaging these maps by using a late fusion (Snoek, Worring and Smeulders, 2005). All maps were equally contributing to the final estimation. In this paper, two scales are used and a different label dependent adaptive weighting scheme is proposed to fuse these maps (eq. 3).

$$p(l) = \alpha(l)p_1(l) + (1 - \alpha(l))p_2(l) \qquad (3)$$

where $p_1(l)$ is the probability map of scale 1 for label *l*, $p_2(l)$ the probability map of scale 2 for label l, $p(l)$ is the final probability map for label l and $\alpha(l) \in [0,1]$ is the mixing coefficient for label l optimized using a gradient descent technique. For this optimization, we minimized the segmentation error *(i.e., maximizing DICE coefficient)*. For this process, since the training dataset was very small (N=5) we used the validation dataset using a leave-one-out approach to avoid bias in the method comparison section. Specifically, we performed 20 estimation rounds (removing one case each time). Details are given in the experiments section.

**Label regularization**

As a consequence of the reduced size of the template library, the segmentations produced by the labeling method may present slightly irregular edges. To reduce this effect we propose a post-processing step to regularize the label maps by convolving them with a smoothing kernel of 5x5x5 voxels. Then, the label with the maximum weight is assigned to each voxel. This penalizes narrow structures being coherent with the layered definition of the cerebellum lobules. To prevent small lobules from being eliminated, the original (not convolved) maps are combined with the corresponding smoothed one using a mixing parameter $\beta$:

$$p_f(l) = p(l) + \beta(l) \cdot H \otimes p(l) \qquad\qquad (4)$$

where $p_f(l)$ represents the final weight of label $l$, $H$ is the convolution kernel and $\beta(l)$ is the regularization mixing coefficient of label $l$ optimized in an orthogonal manner performing a dichotomic search to find the optimal value for each label. We performed the optimization using the validation dataset following a leave-one-out approach as done for the estimation of the alpha coefficients.

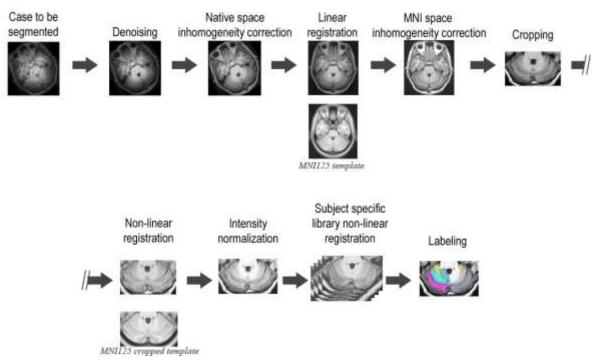Figure 2 presents an overview of the whole pipeline showing the workflow for a new case to be segmented.



*Figure 2: Segmentation pipeline overview.*

**Library extension**

To mitigate the effect of the reduced size of the library of manually segmented templates, an automatic library extension strategy was used. To this purpose, 100 randomly selected unlabeled images from healthy subjects were used. The images belong to the IXI - Information eXtraction from Images (EPSRC GR/S21533/02) dataset which is publicly available at its website (http://biomedic.doc.ic.ac.uk/brain-development).

The library extension strategy consisted of applying the described segmentation pipeline to the new set of images and to add these images and the corresponding segmentations to the library as done in some previous works (Eskildsen et al., 2012; Giraud et al., 2015).

## 3. Experiments and results

Several experiments were performed to estimate optimal method parameters and to compare the performance of the proposed method in terms of accuracy and computational time with other related state-of-the-art approaches.

### *Method parameters*

As OPAL (Giraud et al., 2015) has been developed and evaluated for hippocampus segmentation, an optimization was required to adapt a set of parameters in order to obtain optimal cerebellum segmentation results. To measure the accuracy we used the DICE (Zijdenbos et al., 1994) coefficient for the 13 structures of the cerebellum (left and right labels have been combined into a single one). Results for the following experiments are calculated over the validation set. Note that since only half cerebellum was manually labeled, results were obtained only comparing the segmentation results with the labeled area (discarding the other half cerebellum). In all the experiments we set the patch sizes to 3x3x3 and 9x9x9 voxels, respectively. The restricted search area was set to 7x7x7 voxels. The number of PMs was set to 32 and the number of iterations of OPAL to 4.

### *Library construction*

We mentioned that CERES works using a preprocessed library of manually segmented cases. In this preprocess the images are located in the MNI space by linear registration. Then the images are cropped and registered again to better fit the cerebellum anatomy. This second registration has a significant impact in the segmentation. We tested three different approaches to estimate the impact of the different transformation options. The first one consists on performing a linear registration in the cropped MNI space. In the second approach a non-linear transformation is estimated to register the training templates to the cropped MNI space. Finally, in the third one, we generate a subject specific library by nonlinearly registering all the training templates to the cropped image of the subject in the MNI space. In table 1, the results of these three different options are presented.

As can be noted, the subject specific library construction option was the best one for all the structures. In this experiment, we applied OPAL method without any of the proposed optimizations to measure just the influence of the library construction approach.

*Table 1: Mean and standard deviation DICE results for CERES segmentation over the validation set. Columns show the average values and standard deviation for the lobules and the whole cerebellum. The values are calculated in the native space. Best results marked in bold.*

| Lobule | Linear reg. | Non-linear reg. | Subject specific |
|---|---|---|---|
| AVERAGE | 0.7207 ± 0.0574 | 0.7549 ± 0.0452 | **0.7595** ± 0.0474 |
| Lobule I-II | 0.4022 ± 0.2157 | 0.4165 ± 0.1993 | **0.4226** ± 0.2110 |
| Lobule III | 0.6644 ± 0.2039 | 0.7044 ± 0.1728 | **0.7068** ± 0.1714 |
| Lobule IV | 0.6879 ± 0.1625 | 0.7156 ± 0.1301 | **0.7213** ± 0.1374 |
| Lobule V | 0.7263 ± 0.1424 | 0.7448 ± 0.1315 | **0.7514** ± 0.1376 |
| Lobule VI | 0.8392 ± 0.0437 | 0.8620 ± 0.0322 | **0.8667** ± 0.0322 |
| Lobule Crus I | 0.8678 ± 0.0297 | 0.8937 ± 0.0174 | **0.8970** ± 0.0161 |
| Lobule Crus II | 0.7564 ± 0.0668 | 0.7973 ± 0.0545 | **0.8024** ± 0.0553 |
| Lobule VIIB | 0.6076 ± 0.1354 | 0.6667 ± 0.1107 | **0.6742** ± 0.1107 |
| Lobule VIIIA | 0.7277 ± 0.0705 | 0.7783 ± 0.0691 | **0.7830** ± 0.0692 |
| Lobule VIIIB | 0.8012 ± 0.0541 | 0.8456 ± 0.0383 | **0.8499** ± 0.0358 |
| Lobule IX | 0.8623 ± 0.0383 | 0.8813 ± 0.0322 | **0.8849** ± 0.0313 |
| Lobule X | 0.7053 ± 0.0830 | 0.7529 ± 0.0438 | **0.7540** ± 0.0473 |
| Cerebellum | 0.9274 ± 0.0107 | 0.9346 ± 0.0091 | **0.9358** ± 0.0091 |

### Muli-scale mixing coefficient

As explained before, OPAL obtains the matches for every image patch at different scales and then performs a late fusion. We compared the uniform averaging scheme ($\alpha$=0.5) as done in original OPAL paper with the proposed adaptive method.  Table 2 shows the results for the two compared options. Note that we optimized these coefficients 20 times using leave-one-out cross validation. Results in table 2 show the average dice of the 20 cases using its corresponding set of coefficients (estimated from the 19 remaining cases of the validation set, i.e. excluding the case under study). In Tab. 3, for sake of simplicity we only show the average of the 20 different estimations. In following experiments, the 20 different sets of alpha coefficients will be used for validation to prevent the validation cases to be included in its own segmentation process.

*Table 2: Mean DICE results for CERES segmentation over the validation set. Columns show the average values for the different lobules and the whole cerebellum for the two compared approaches. Best results marked in bold.*

| Lobule | $\alpha$ = 0.5 | $\alpha$ adaptive (LOOCV) |
|---|---|---|
| AVERAGE | 0.7595 | **0.7640** |
| Lobule I-II | 0.4226 | **0.4364** |
| Lobule III | 0.7068 | **0.7237** |
| Lobule IV | 0.7213 | **0.7281** |
| Lobule V | 0.7514 | **0.7517** |
| Lobule VI | 0.8667 | **0.8673** |

| | | | |
|---|---|---|---|
| Lobule Crus I | 0.8970 | | **0.9000** |
| Lobule Crus II | 0.8024 | | **0.8034** |
| Lobule VIIB | 0.6742 | | **0.6756** |
| Lobule VIIIA | 0.7830 | | **0.7848** |
| Lobule VIIIB | 0.8499 | | **0.8506** |
| Lobule IX | 0.8849 | | **0.8855** |
| Lobule X | 0.7540 | | **0.7604** |
| Cerebellum | 0.9358 | | **0.9370** |

*Table 3: Values for the adaptive $\alpha$ coefficient optimized using a gradient descent technique.*

| Label | $\alpha$ coefficient | Label | $\alpha$ coefficient |
|---|---|---|---|
| Lobule I-II | 0.82 | Lobule VIIB | 0.59 |
| Lobule III | 0.81 | Lobule VIIIA | 0.59 |
| Lobule IV | 0.71 | Lobule VIIIB | 0.61 |
| Lobule V | 0.59 | Lobule IX | 0.60 |
| Lobule VI | 0.50 | Lobule X | 0.79 |
| Lobule Crus I | 0.81 | WM | 0.51 |
| Lobule Crus II | 0.60 | Background | 0.82 |

**Impact of label regularization**

We performed a similar experiment to measure the effect of the label regularization on the segmentation results. Table 4 shows the results of the proposed regularization method computed over the 20 subjects within the validation dataset. Note that we optimized these coefficients 20 times using leave-one-out cross validation as previously described and the final DICE coefficients shown in table 4 were computed using the corresponding parameter set from the leave-one-out estimation. Also, in the following experiments, the 20 different sets of beta coefficients will be used to prevent the validation cases to be included in its own segmentation process.

Despite the fact that the global improvement is limited, an important increase in lobule I-II, which is a difficult structure to segment given its high variability within subjects, can be seen due to the regularization. The optimal values of α coefficient were 18 for lobule I-II and 4 for the rest of structures including the background label. To more specifically evaluate the impact on the label contour, we calculated the Hausdorff distance (Beauchemin et al., 1998) for each label. Despite the differences not being significant, regularization tends to produce lower distances.

*Table 4: Mean DICE and Hausdorff distance results for CERES segmentation over the validation set. Columns show the average values for the lobules and the whole cerebellum. The values are calculated in the native space. Best results are marked in bold.*

| Structure | No regularization | | Regularization | |
|---|---|---|---|---|
| | DICE | Hausdorff | DICE | Hausdorff |

| | | | | |
|---|---|---|---|---|
| **AVERAGE** | 0.7640 | 6.3992 | **0.7702** | **6.3302** |
| **Lobule I-II** | 0.4364 | 3.2214 | **0.5056** | **2.8571** |
| **Lobule III** | **0.7237** | 6.2210 | 0.7157 | **6.0854** |
| **Lobule IV** | 0.7281 | 5.5987 | **0.7320** | **5.4904** |
| **Lobule V** | 0.7517 | 6.7155 | **0.7535** | **6.7022** |
| **Lobule VI** | 0.8673 | 6.1831 | **0.8678** | **6.1557** |
| **Lobule Crus I** | 0.9000 | 6.5719 | **0.9005** | **6.5583** |
| **Lobule Crus II** | 0.8034 | 10.9254 | **0.8043** | **10.8393** |
| **Lobule VIIB** | 0.6756 | **9.8729** | **0.6778** | 9.9934 |
| **Lobule VIIIA** | 0.7848 | 8.4833 | **0.7862** | **8.3133** |
| **Lobule VIIIB** | 0.8506 | 4.7480 | **0.8531** | **4.7222** |
| **Lobule IX** | 0.8855 | **5.0390** | **0.8858** | 5.0499 |
| **Lobule X** | **0.7604** | 3.2097 | 0.7603 | **3.1952** |
| **Cerebellum** | 0.9370 | - | **0.9372** | - |

### *Impact of extended library size*

In this experiment we evaluated the impact of the number of subjects from the extended library used to perform the labeling step. The extended library included the original 10 manually labeled templates and 100 automatically segmented images from IXI dataset. From the library, the $N$ most similar cases (in the mean squared error sense) to the target are selected. As can be noticed in Figure 3 the plateau is reached around 50 cases. Note that we have a noticeable oscillation for values higher than 50. As explained before, PatchMatch algorithm involves a random search step that is performed iteratively. In contrast to an exhaustive search PatchMatch converges faster because it doesn't explores the whole library. Thus we obtain an approximated nearest neighbor field. This implies that libraries with different sizes will have a different set of visited patches even if the majority of the templates are the same. For this reason, library sizes larger than the optimal (50) present oscillation around the optimal result. The use of this library extension enabled to improve the mean DICE coefficient from 0.7702 ($N$=10) to 0.7729 ($N$=50). Therefore, we set $N$=50 as the default parameter of the method.
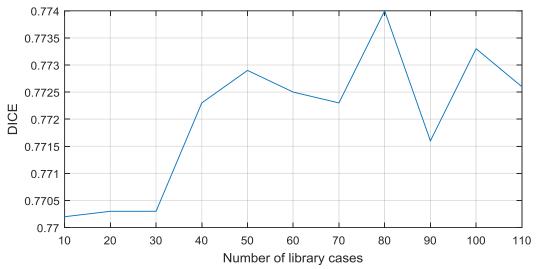
*Figure 3: Evolution of DICE coefficient by number of cases from the extended library used for segmentation over the validation set.*

### Method comparison

The proposed method (with all the optimizations and the extended library) was compared with 3 state-of-the-art methods: SUIT (Diedrichsen, 2006), MAGeT (Park et al., 2014; Chakravarty et al., 2013) and RASCAL (Weier et al., 2014). The different compared methods were applied to segment the validation dataset to evaluate their performance. In Table 5 CERES method is compared with MAGeT and RASCAL methods for all 12 structures and the whole cerebellum. As can be noticed, CERES outperformed both method overall (only MAGeT had a slightly better result for lobule VI).

*Table 5: Mean DICE coefficient and standard deviation values for MAGeT, RASCAL and CERES. Intra-rater accuracy for the manual segmentations from the validation set is also shown for evaluation purposes. Best results are marked in bold (significant differences (p<0.05) are marked with * for MAGeT and † for RASCAL comparison).*

| Structure | MAGeT | RASCAL | CERES | Intra-rater manual |
|---|---|---|---|---|
| *AVERAGE* | 0.7320 ± 0.0568 | 0.6890 ± 0.0524 | **0.7729** ± 0.0427*†* | 0.833 |
| *Lobule I-II* | 0.3960 ± 0.1424 | 0.3260 ± 0.2178 | **0.5201** ± 0.1555**†* | 0.639 |
| *Lobule III* | 0.6800 ± 0.1741 | 0.6379 ± 0.2165 | **0.7213** ± 0.1572 | 0.751 |
| *Lobule IV* | 0.6980 ± 0.1440 | 0.6627 ± 0.1611 | **0.7271** ± 0.1346 | 0.818 |
| *Lobule V* | 0.7320 ± 0.1398 | 0.6666 ± 0.1560 | **0.7561** ± 0.1332*†* | 0.881 |
| *Lobule VI* | **0.8710** ± 0.0359 | 0.7969 ± 0.0523 | 0.8695 ± 0.0316*†* | 0.912 |
| *Lobule Crus I* | 0.8870 ± 0.0257 | 0.8383 ± 0.0351 | **0.9007** ± 0.0152*†* | 0.904 |
| *Lobule Crus II* | 0.7780 ± 0.0679 | 0.7340 ± 0.0667 | **0.8096** ± 0.0569*†* | 0.900 |
| *Lobule VIIB* | 0.5990 ± 0.1487 | 0.5820 ± 0.1137 | **0.6850** ± 0.1205**†* | 0.863 |
| *Lobule VIIIA* | 0.7300 ± 0.0934 | 0.6757 ± 0.1426 | **0.7926** ± 0.0759**†* | 0.860 |

| | | | | |
|---|---|---|---|---|
| **Lobule VIIIB** | 0.7970 ± 0.0607 | 0.7783 ± 0.0931 | **0.8533** ± 0.0390*† | 0.833 |
| **Lobule IX** | 0.8560 ± 0.0384 | 0.8460 ± 0.0545 | **0.8849** ± 0.0327† | 0.874 |
| **Lobule X** | 0.7540 ± 0.0490 | 0.7237 ± 0.0680 | **0.7548** ± 0.0469* | 0.760 |
| **Cerebellum** | 0.9250 ± 0.0094 | 0.9349 ± 0.0089 | **0.9377** ± 0.0090† | 0.941 |

As SUIT works with a different set of labels, the output segmentations of MAGeT, RASCAL and CERES were adapted by fusing lobules I-II, III and IV for a meaningful comparison as done in Park et al. (2014). Lobules Crus I and IX were not taken into account due to differences in their definition. We processed the validation images using SUIT as described in http://www.icn.ucl.ac.uk/motorcontrol/imaging/suit_function.htm without applying any manual correction. In Table 6, DICE coefficient values for the different compared methods are shown. CERES provided the best results overall (MAGeT results were better for lobule I-IV, Vi and X although differences were not significant). In Figure 4, some example images from the different method segmentations are shown.

Table 6: Mean and standard deviation for DICE coefficient values for SUIT, MAGeT, RASCAL and CERES for the SUIT segmentation protocol. Best results are marked in bold (significant differences (p<0.05) are marked with * for SUIT, † for MAGeT and ‡ for RASCAL comparison).

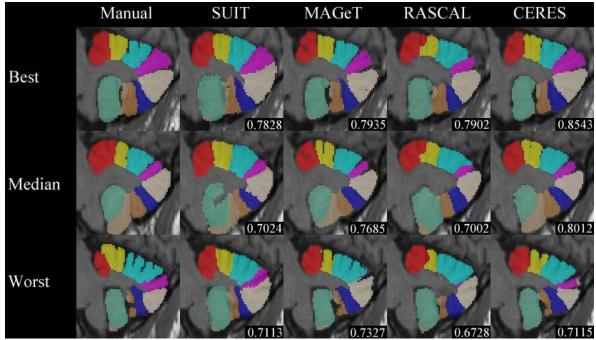| Structure | SUIT | MAGeT | RASCAL | CERES |
|---|---|---|---|---|
| **AVERAGE** | 0.7097 ± 0.0689 | 0.7639 ± 0.0792 | 0.7140 ± 0.0487 | **0.7888** ± 0.0409 *‡ |
| **Lobule I-IV** | 0.7435 ± 0.0880 | **0.8055 ± 0.0964** | 0.7703 ± 0.1107 | 0.7898 ± 0.1021 * |
| **Lobule V** | 0.6598 ± 0.1026 | 0.7429 ± 0.1369 | 0.6730 ± 0.1560 | **0.7561** ± 0.1332 *‡ |
| **Lobule VI** | 0.7800 ± 0.0543 | **0.8762 ± 0.0365** | 0.7994 ± 0.0523 | 0.8695 ± 0.0316 *‡ |
| **Lobule Crus II** | 0.7430 ± 0.0631 | 0.7787 ± 0.0678 | 0.7300 ± 0.0667 | **0.8096** ± 0.0569 *‡ |
| **Lobule VIIB** | 0.5701 ± 0.1572 | 0.6013 ± 0.1476 | 0.5761 ± 0.1137 | **0.6850** ± 0.1205 *‡ |
| **Lobule VIIIA** | 0.7134 ± 0.0996 | 0.7330 ± 0.0928 | 0.6701 ± 0.1426 | **0.7926** ± 0.0759 *†‡ |
| **Lobule VIIIB** | 0.7721 ± 0.0596 | 0.8012 ± 0.0607 | 0.7654 ± 0.0931 | **0.8533** ± 0.0390 *†‡ |
| **Lobule X** | 0.6955 ± 0.0512 | **0.7721 ± 0.0475** | 0.7275 ± 0.0680 | 0.7548 ± 0.0469 * |

*Figure 4: Manual, SUIT, MAGeT, RASCAL and CERES segmentation examples over the cases with the best, median and worst DICE for CERES. Corresponding DICE is provided for each segmentation. Note how SUIT tends to produce oversegmentation and always fills the fissures between lobules. Also note the irregular boundary between lobules I-IV(red) and V(yellow) produced by RASCAL in all three cases. Finally, in the worst case, it can be seen that lobule VIIIA(purple) in MAGeT and RASCAL segmentations is barely visible.*

### Execution time

One of the main advantages of our method is its efficiency as it produces competitive results in terms of accuracy with a reduced execution time. The slowest method in the comparison was MAGeT (approximately 4 hours per case) due to the need of performing many non-linear registrations. RASCAL takes around 110 minutes to provide the segmentation result. Finally, SUIT (without manual correction) and CERES require around 5 minutes to produce the segmentation. Figure 5 shows the computational cost of the different parts of the proposed method. As can be noticed the most time consuming part is the non-linear registration estimation. Note also the segmentation part took only 1 minute thanks to the efficiency of the OPAL method, while the preprocessing remains the most time consuming step of the proposed pipeline.
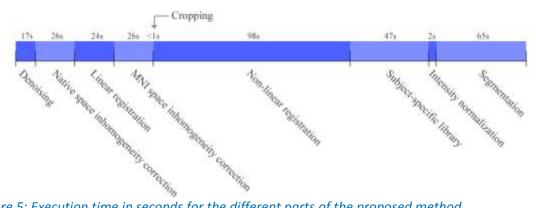
*Figure 5: Execution time in seconds for the different parts of the proposed method.*

## 4. Discussion

We have presented a new pipeline for cerebellum lobule segmentation called CERES that works in a fully automatic manner and that is able to provide accurate results in a low computation time. We have shown that CERES produces state-of-the-art results outperforming current cerebellum segmentation methods such as SUIT, MAGeT and RASCAL. Main contributions of the presented method are: 1) the use of an efficient custom made non-linear library for each case to be segmented 2) an adaptive multi-scale implementation and 3) a new label regularization step.

In terms of accuracy, CERES obtained the best results for most of the labels. Each of the improvements proposed in the method allowed to consistently leverage the results being the custom library construction the most significant one (0.7207 to 0.7595). The adaptive multi-scale mixing and the label regularization steps allowed to further boost the results to obtain 0.7729 final DICE coefficient. More interestingly, CERES almost reaches intra-rater accuracy for the whole cerebellum segmentation (CERES = 0.9377 *vs.* human = 0.941) (Park et al., 2014). It is worth noting that the high accuracy obtained by CERES was achieved using only 5 manually labeled templates. It is expected that results with larger manually labeled template libraries will provided significantly better results by covering a larger variability of cerebellum shapes and conditions (including pathological ones).

From the efficiency point of view, CERES was found to produce the segmentation results in a very short time (5 minutes per case) being the most efficient of the compared methods. This can be explained by two main reasons. First, multi-atlas label fusion techniques usually require the processing of multiple non-linear registrations which is quite time consuming. CERES only requires a single non-linear registration since it applies the *N* non-linear mapping from a set of pre-computed transformation to the MNI space. Furthermore, this single non-linear registration being low-dimensional and applied to a cropped volume is much faster estimated that high dimensional registration typically used in multi-atlas methods. Second, one of the most computationally demanding steps in non-local multi-atlas label fusion methods is the patch matching process. In

CERES, this time is highly reduced by using OPAL making the patch matching process no longer the bottleneck of the segmentation procedure.

Another important feature of CERES is its adaptability. As it uses a library of manually segmented templates as a knowledge base it can learn new anatomies by adding new cases to the library. In our experiments, the main limitation of CERES is the small size of the manually labeled template library (only 5 cases). In previous multi-atlas label fusion approaches it has been shown that a library of at least around 20 cases is required in order to achieve optimal segmentation results. Therefore, it is expected that CERES performance will be improved when using a larger library. This will be done by adding more manually segmented cases as long as they become available.

Finally, in order to make CERES pipeline fully and freely accessible to the scientific community, we have integrated this pipeline as a part of our online volBrain platform ([http://volbrain.upv.es](http://volbrain.upv.es)). At the time of writing this paper, CERES pipeline has been already applied to more than 3000 cases (in around 5 months) with a failure rate lower than 1%.

**Acknowledgements**

**References**

Aljabar, P., Heckermann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *Neuroimage* 46(3), 726 - 738.

Asman, A.J., Landman, B.A., 2012. Muti-Atlas Segmentation using Non-Local STAPLE. *MICCAI workshop on Multi-Atlas Lebling*, Nice, 87 - 90.

Avants, B.B., Tustison, N., Song, G., 2009. Advanced normalization tools (ANTS). *Insight Journal*.

Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B., 2009. PatchMatch: a randomized correspondence algorithm for structural image editing. *ACM Trans Graph* 28(3).

Beauchemin M., Thomson K.P.B., Edwards G., 1998. On the Hausdorff distance used for the evaluation of segmentation results. *Canadian Journal of Remote Sensing* 24(1), 3 - 8.

Bogovic, J.A., Bazin, P.L., Ying, S.H., Prince, J.L., 2013. Automated segmentation of the cerebellar lobules using boundary specific classification and evolution. *Inf Process Med Imaging* 23, 62 - 73.

Chakravarty, M.M., Steadman, P., Eede, M.C., Calcott, R.D., Gu, V., Shaw, P., Raznahan, A., Collins, D.L., Lerch, J.P., 2013. Performing label-fusion-based segmentation using multiple automatically generated templates. *Human brain mapping*, 10(34), 2635 - 2654.

Compston, A., Coles, A., 2008. Multiple sclerosis. *Lancet* 372, 1502 - 1517.

Collins, D.L., Neelin, P., Peters, T. M., Evans, A. C., 1994. Automatic 3D Inter-Subject Registration of MR Volumetric Data in Standardized Talairach Space. *Journal of Computer Assisted Tomography,* 18(2) 192 - 205.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based Segmentation using Expert Priors: Application to Hippocampus and Ventricle Segmentation. *NeuroImage* 54, 940 - 954.

Davie, C.A., Barker, G.J., Webb, S., Tofts, P.S., Thompson, A.J., Harding, A.E., McDonald, W.I., Miller, D.H., 1995. Persistent functional deficit in multiple sclerosis and autosomal dominant cerebellar ataxia is associated with axon loss. *Brain* 6, 1583 - 1592.

Diedrichsen, J., 2006. A spatially unbiased atlas template of the human cerebellum. *NeuroImage* 33, 127 - 138.

Eskildsen, S.F., Coupé, P., Fonov, V., Manjón, J.V., Leung, K.K., Guizard, N., Wassef, S.N., Østergaard, L.R., Collins, D.L., Initiative, A.D.N., 2012. BEaST: brain extraction based on nonlocal segmentation technique. *Neuroimage* 59(3):, 2362 - 2373.

Giraud, R., Ta, V.-T., Papadakis, N., Manjón, J.V., Collins, D.L., Coupé, P., Initiative, A.D.N., 2015. An Optimized PatchMatch for multi-scale and multi-feature label fusion. *Neuroimage* In press.

James, A.C., James, S., Smith, D.M., Javaloyes, A., 2004. Cerebellar, prefrontal cortex, and thalamic volumes over two time points in adolescent-onset schizophrenia. *Am J Psychiatry* 161, 1023 - 1029.

Kase, C.S., Norrving, B., Levine, S.R., Babikian, V.L., Chodosh, E.H., Wolf, P.A., Welch, K.M., 1993. Cerebellar infarction. Clinical and anatomic observations in 66 cases. *Stroke* 24, 76 - 83.

Klockgether, T., 2008. The clinical diagnosis of autosomal dominant. *Cerebellum* 7, 101 - 105.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M., 2010. Adaptive non-local means denoising of MR images with spatially varying noise levels. *J Magn Reson Imaging* 31(1), 192 - 203.

Manjón, J.V., Eskildsen, S.F., Coupé, P., Romero, J.E., Collins, D.L., Robles, M., 2014. Nonlocal Intracranial Cavity Extraction. *IJBI, ID 820205*.

Manjón, J.V., Tohka, J., Robles, M., 2010. Improved Estimates of Partial Volume Coefficients from Noisy Brain MRI Using Spatial Context. *Neuroimage* 53, 480 - 490.

Manto, M., Gruol, D., Schmahmann, J., Koibuchi, N., Rossi, F., 2013. Handbook of the Cerebellum and Cerebellar Disorders. *New York: Springer*.

Nenadic, I., Sauer, H., Gaser, C., 2010. Distinct pattern of brain structural deficits in subsyndromes of schizophrenia delineated by psychopathology. *Neuroimage* 49:, 1153 - 1160.

Okugawa, G., Sedvall, G., Nordström, M., Andreasen, N., Pierson, R., Magnotta, V., Agartz, I., 2002. Selective reduction of the posterior superior vermis in men with chronic schizophrenia. *Schizophr Res* 55, 61 - 67.

Park, M.T.M., Pipitone, J., Baer, L.H., Winterburn, J.L., Shah, Y., Chavez, S., Schira, M.M., Labaugh, N.J., Lerch, J.P., Vineskos, A.N., Chakravarty, M.M., 2014. Derivation of high-resolutionMRI atlases of the human cerebellumat 3 T and segmentation using multiple automatically generated templates. *NeuroImage* 95, 217 - 231.

Rousseau, F., Habas, P.A., Studholme, C., 2011. A supervised patch-based approach for human brain labeling. *IEEE Transactions on Medical Imaging* 30(10), 1852 - 1862.

Snoek, C.G., Worring, M., Smeulders, A.W., 2005. Early versus late fusion in semantic video analysis. *ACM International Conference on Multimedia* 399 - 402.

Stoodley, C.J., 2012. The cerebellum and cognition: evidence from functional imaging studies. *Cerebellum* 11(2), 352 - 356.

Ta, V.T., Giraud, R., Collins, D.L., Coupé, P., 2014. Optimized patchMatch for near real time and accurate label fusion. *Med Image Comput Comput Assist Interv* 17, 105 - 112.

Thomann, P.A., Schläfer, C., Seidi, U., Santos, V.D., Essig, M., Schröder, J., 2008. The cerebellum in mild cognitive impairment and Alzheimer's disease - a structural MRI study. *J Psychiatr Res* 42, 1198 - 1202.

Timmann, D., Drepper, J., Frings, M., Maschke, M., Richter, S., Grewing, M., Kolb, F.P., 2010. The human cerebellum contributes to motor, emotional and cognitive associative learning. A review. *Cortex* 46, 845 - 857.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* 29(6), 1310 - 1320.

Wang, H., Avants, B., Yushkevich, P.A., 2012. Grand Challenge on Multi-Atlas Segmentation: A Combined Joint Label Fusion and Corrective Learning Approach. *MICCAI workshop on Multi-Atlas Lebling*, Nice, 91 - 94.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 600 - 612.

Weier, K., Beck, A., Magon, S., Amann, M., Naegelin, Y., Penner, I.K., Thürling, M., Aurich, V., Defruss, T., Radue, E.-W., Stippich, C., Kappos, L., Timmann, D., Sprenger, T., 2012. Evaluation of a new approach for semi-automatic segmentation of the cerebellum in patients with multiple sclerosis. *Journal of Neurology* 259, 2673 - 2680.

Weier, K., Fonov, V., Lavoie, K., Doyon, J., Collins, D.L., 2014. Rapid Automatic Segmentation of the Human Cerebellum and its Lobules (RASCAL) - Implementation and Application of the Patch-based Label-fusion Technique With a Template Library to Segment the Human Cerebellum. *Human Brain Mapping* 5026 - 5039.

Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D., Initiative, A.D.N., 2010. LEAP: Learning embeddings for atlas propagation. *Neuroimage* 49(2), 1316 - 1325.

Yang, Z., Ye, C., Bogovic, J.A., Carass, A., Jedynak, B.M., Ying, S.H., Prince, L.J., 2016. Automated cerebellar lobule segmentation with application to cerebellar structural analysis in cerebellar disease. *Neuroimage* 127, 435 - 444.

Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palme, A.C., 1994. Morphometric analysis of white matter lesions in MR images: method and validation. *IEE Trans Med Imaging* 13, 716 - 724.