An Optimized PatchMatch for multi-scale and multi-feature label fusion

Rémi Giraud, Vinh-Thong Ta, Nicolas Papadakis, Jose Vicente Manjon, D Louis Collins, Pierrick Coupé

An Optimized PatchMatch for Multi-scale and Multi-feature Label Fusion

Rémi Giraud[a,b,c,d,e,*], Vinh-Thong Ta[a,b,e], Nicolas Papadakis[c,d], José V. Manjón[f], D. Louis Collins[g], Pierrick Coupé[a,b], and the Alzheimer's Disease Neuroimaging Initiative [*]

[a] *Univ. Bordeaux, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France.*
[b] *CNRS, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France.*
[c] *Univ. Bordeaux, IMB, UMR 5251, F-33400 Talence, France.*
[d] *CNRS, IMB, UMR 5251, F-33400 Talence, France.*
[e] *Bordeaux INP, LaBRI, UMR 5800, PICTURA, F-33600 Pessac, France.*
[f] *Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.*
[g] *McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada.*

Abstract

Automatic segmentation methods are important tools for quantitative analysis of Magnetic Resonance Images (MRI). Recently, patch-based label fusion approaches have demonstrated state-of-the-art segmentation accuracy. In this paper, we introduce a new patch-based label fusion framework to perform segmentation of anatomical structures. The proposed approach uses an Optimized PAtchMatch Label fusion (OPAL) strategy that drastically reduces the computation time required for the search of similar patches. The reduced computation time of OPAL opens the way for new strategies and facilitates processing on large databases. In this paper, we investigate new perspectives offered by OPAL, by introducing a new multi-scale and multi-feature framework. During our validation on hippocampus segmentation we use two datasets: young

[*]email: remi.giraud@labri.fr, tel: +33540006937, fax: +33540006669

[*]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.use.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`.

adults in the ICBM cohort and elderly adults in the EADC-ADNI dataset. For both, OPAL is compared to state-of-the-art methods. Results show that OPAL obtained the highest median Dice coefficient (89.9% for ICBM and 90.1% for EADC-ADNI). Moreover, in both cases, OPAL produced a segmentation accuracy similar to inter-expert variability. On the EADC-ADNI dataset, we compare the hippocampal volumes obtained by manual and automatic segmentation. The volumes appear to be highly correlated that enables to perform more accurate separation of pathological populations.

Keywords: Patch Matching, Segmentation, Late Fusion, Hippocampus, Patch-Based Method.

1. Introduction

Magnetic Resonance Imaging (MRI) has become an essential tool in medical analysis, especially in the study of the human brain. The segmentation of MRI brain structures is a necessary step for many clinical applications. The manual segmentation of structures in MRI by clinical experts is still considered as the gold standard. However, manual labeling is a highly tedious and very time consuming task. Moreover, the manually generated segmentations are subject to inter- and intra-rater variability. Therefore, designing fast, accurate and reliable automatic segmentation methods is a challenging work in quantitative MRI analysis.

In the past decade, several paradigms were proposed to automatically perform brain segmentation. First, atlas-based methods involving nonlinear registration of a labeled atlas to the subject were proposed [1, 2]. Once the atlas is matched to the subject image, the segmentation is achieved by warping the atlas labels to the target image space. Such atlas-based methods have been widely used due to their robustness and the ease of integration of expert priors. However, atlas-based methods may not sufficiently capture inter-subject variability due to the one-to-one mapping assumption between the atlas and the subject anatomy. Consequently, atlas-based methods are subject to registration errors

since in general such mapping does not exist.

In order to minimize registration errors, template warping techniques based on a training library of manually labeled templates were introduced. The simplest method based on a library of training templates is the best-template approach [3]. The main idea is to reduce the anatomical distance between a selected template and the subject to be segmented in order to improve registration accuracy. First, the most similar template is selected in the training library. Then, this template is nonlinearly registered to the subject. Finally, the estimated nonlinear transformation is applied to the manually segmented labels in the selected template to obtain the final segmentation. While the selection of the most similar template compared to an *a priori* fixed atlas may improve segmentation results, the best template strategy is still subject to registration errors and leads to sub-optimal results.

A significant improvement has been obtained with the introduction of multi-template approaches. Such methods merge information from several similar training templates instead of using a single template to achieve better segmentation. In such methods, the registration errors resulting from inter-subject variability are considered as a random variable, thus reducing segmentation error by using several atlases [4, 5]. Since its introduction, many approaches have been proposed to improve the label fusion step, such as preselection of most similar template following by majority voting [6, 7, 8], intensity models [9, 10], fusion techniques with local weighted label fusion [11, 12, 13] or systematic bias correction using a learning-based method [14]. Multi-templates matching approaches demonstrated competitive segmentation accuracy at the expense of an important computational burden resulting from multiple nonlinear registrations, *i.e.*, up to several hours.

Recently, a nonlocal patch-based label fusion (PBL) method [15] has been proposed for reducing the computational burden of multi-templates based methods. Instead of performing multiple nonlinear registrations, the PBL method relies on the comparison of patches (centered neighborhood around a voxel) which only requires an affine alignment of the subject and the training tem-

plates. The patch comparisons performed between the current image patch and training patches, are used to assign a weight to the manual labels according to patch similarity. The search for similar training patches is based on a nonlocal strategy in order to better capture registration inaccuracies and to efficiently handle the inter-subject variability. PBL overcomes the one-to-one mapping assumption of multi-template warping methods thanks to a well-defined one-to-many mapping model. Finally, the PBL approach produces state-of-the-art segmentation accuracy with limited computation time, *i.e.*, several minutes.

Since its introduction, the PBL approach has been intensively studied and many improvements have been proposed. First, PBL can be combined with other methods such as multi-template warping [16], active appearance models [17] or level sets [18]. Moreover, other improvements have been proposed using multi-resolution framework [19], discriminative dictionary learning and sparse coding [20], or generative probability models [21]. However, PBL still suffers from several limitations. First, the search for similar patches is still computationally expensive. Although preselection of templates and patches [15] or multi-scale strategies [19] have been proposed, an important amount of computation remains dedicated to the search for similar patches in the training library. Secondly, the template preselection step can prevent finding the most similar patches existing in the library. By selecting training templates according to a global similarity measure between the subject and the template, the template preselection step is likely to remove relevant parts of the training library, possibly leading to sub-optimal results. Finally, in PBL, patch comparisons are performed between the current patch and training patches. The relevance of the match is then weighted depending on the similarity between the two patches. However, weights are assigned to a large number of training patches including many dissimilar patches. Beyond inefficient computations dedicated to estimate negligible weights, these dissimilar patches can decrease the segmentation accuracy [20]. Sparsity-based methods tend to limit this issue but suffer from an important computational burden [20, 21].

In this paper, we first introduce a new Optimized PAtchMatch for Label

fusion (OPAL) to address the limitations of previous PBL approaches in terms of computation time and search strategy of similar patches. The OPAL method is able to find, in significantly less computations, similar patches over the entire training library without template or patch preselection. Originally, the PatchMatch (PM) [22] algorithm was introduced to efficiently find patch correspondences between two 2D images. For each patch within the first image, an approximate nearest neighbor (ANN) is found within the second image. The algorithm is based on a cooperative and randomized strategy resulting in very low computation time, enabling near real-time processing. PM has been applied to medical imaging for super-resolution of cardiac MRI [23], but most PM applications concern 2D image editing problems. In this work, we investigate the use of PM for anatomical structures segmentation using multi-templates training library. Thanks to our Optimized PM (OPM) algorithm, OPAL produces segmentations in a few seconds compared to previous PBL methods. Beyond computation time efficiency, OPAL complexity only depends on the size of the area to be processed within the subject. Consequently, our method does not require any preselection, since the search of most similar patches is achieved over the entire training library. Without training template or patch preselection, similar patches can be found within the whole template library leading to higher segmentation accuracy.

The drastically reduced computation time of OPAL opens the way for new strategies and efficient processing of very large databases. In this paper, we investigate new perspectives offered by OPAL by introducing a new multi-scale and multi-feature framework. In our approach, several scales and features are analyzed at the same time before performing the label fusion. First, the OPM is achieved with different patch sizes on each feature. Then, we perform a late fusion of these independent estimators, each one providing different information on structure characteristics. The description of the structures indeed depends on the considered patch size or the image features used. By using multi-scale and multi-feature searches, the diversity of selected matches is improved which increases the segmentation accuracy.

The main contributions of this work are: (i) An adaptation of the PM algorithm to label fusion for anatomical structure segmentation in 3D MRI, including acceleration techniques such as constrained initialization, parallel processing and optimized distance computation; (ii) A novel late fusion strategy of multi-scale and multi-feature estimator maps; (iii) An extensive OPAL validation on hippocampus segmentation on two datasets with comparison to state-of-the-art methods in terms of computation time and segmentation accuracy; and (iv) A comparison of the ability to separate populations, based on hippocampal volumes obtained with manual and automatic segmentation.

2. Methods

2.1. Fast Nearest Neighbor Matching

In the PBL method, the first step consists in finding, for each patch of the subject to segment, relevant matches, *i.e.*, approximate nearest neighbors (ANN), within the training template library. The two main issues of this method are the relevance of the selected patches and the computational burden dedicated to this search. In this work, we propose a fast patch-based nearest neighbor matching algorithm to find highly similar patches, thus addressing the computational costs usually associated with classic PBL techniques.

2.1.1. The PatchMatch Algorithm

The original PM algorithm [22] is a fast and efficient approach that computes patch correspondences (matches) between two 2D images (*e.g.* A & B). The key point of this method is that good matches can be propagated to the adjacent patches within an image. This propagation, combined with random matches, leads to a very fast convergence with limited computational burden. The core of the algorithm is based on three steps: initialization, propagation, and random search. The initialization consists in randomly associating each patch of A with a corresponding patch in B, in order to obtain an initial ANN field. The two following steps are then performed iteratively in order to improve the ANN

field. The propagation step uses the assumption that when a patch p centered on $\mathbf{x_i} = (x, y) \in A$ matches well with a patch q centered on $\mathbf{x_j} \in B$, then the adjacent patches of $p \in A$ should match well with the adjacent patches of $q \in B$. The iterative process follows a scan order (from left to right, top to bottom) on even iterations and is reversed on odd iterations. Therefore, only recently processed pixels are selected to propagate good matches to their neighbors. For example, on even iterations, for a patch located at $\mathbf{x_i} = (x, y) \in A$, only the neighboring patches centered on $(x-1, y)$ and $(x, y-1)$ are considered during the propagation step. Let $\mathbf{x'_j} \in B$ be the match of the patch centered on position $(x-1, y) \in A$. The candidate to improve p correspondence is the patch centered on $\mathbf{x'_j} + (1, 0) \in B$.

Next, the random search step consists of a random sampling around the current ANN to escape from local minima. The candidates are randomly selected within an exponentially decreasing search window centered on $\mathbf{x_j}$. The propagation of good matches within the iterative process combined with random search, provides a very fast convergence of the algorithm in practice.

2.1.2. Optimized PatchMatch Algorithm

In contrast to [22] where two 2D images are considered, OPAL finds the patch correspondences between a 3D image S and a library of n 3D templates $T = \{T_1, \ldots, T_n\}$. One advantage of the PM algorithm is that its complexity only depends on the size of image A to process and not on the size of the compared image B, *i.e.*, T in the OPAL case. This important fact enables OPAL to consider the entire image library T without any template preselection step at constant complexity in time. Moreover, for each patch in S, OPAL computes the best k-ANN matches in T and not only one match as done in [22].

The OPAL algorithm is explained in detail in the next section and Figure 1 proposes a schematic overview. To clearly illustrate our Optimized PatchMatch (OPM) key steps, in Figure 1, only three templates are considered as template library T, two iterations are performed and 3D MRI volumes are displayed in 2D.

As in the original paper, the metric used to compare the distance between a patch centered on $\mathbf{x_i} \in A$ and a patch centered on $\mathbf{x_j} \in B$, is a sum of squared differences (SSD),

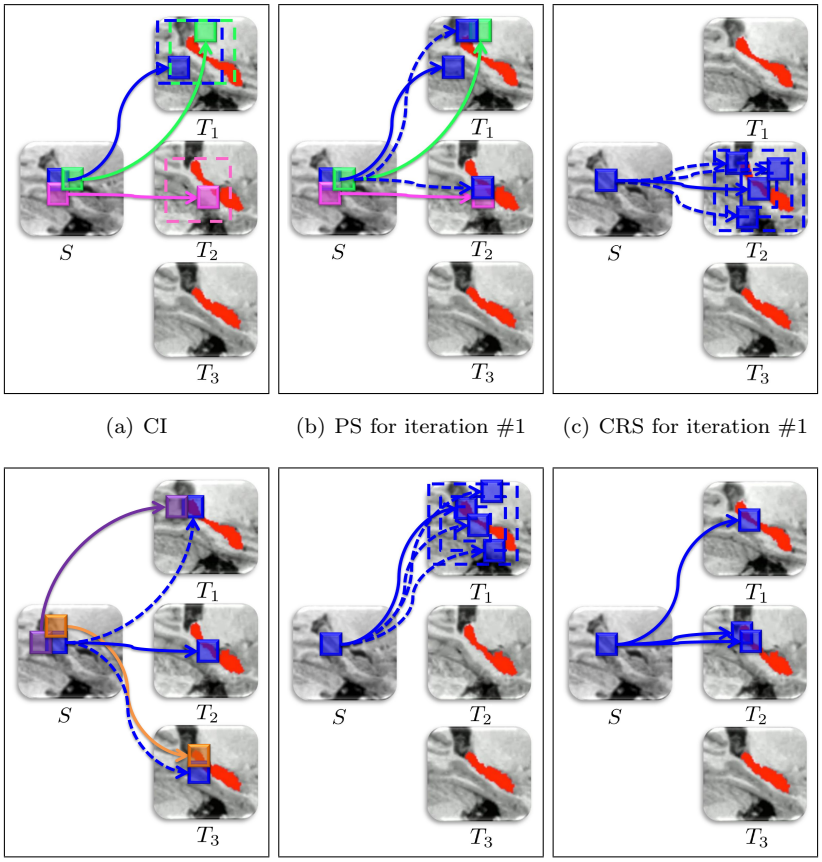
$$dist(\mathbf{x_i}, \mathbf{x_j}) = \sum_{\sigma \in \Omega_s} (A(\mathbf{x_i} + \sigma) - B(\mathbf{x_j} + \sigma))^2, \quad (1)$$

where Ω_s is the index coordinate set of the $s \times s$ 2D patch, centered on $(0, 0)$, considering s as the patch size.

2.1.3. Constrained Initialization

In the PM original paper [22], the initialization consists in assigning, for each patch located at $(x, y) \in A$, a random correspondence which can be located everywhere at $(x', y') \in B$. In the case of multi-templates method based on 3D MRI, the natural extension of this initialization step is to assign, for each patch of the 3D image of the subject to segment S located at $\mathbf{x_i} = (x, y, z) \in S$, a random patch correspondence located at $\mathbf{x_j} = \{(x', y', z'), t\}$ where $t \in \{1, \ldots, n\}$ is the index of the template T_t within the template library T. However, as we deal with linearly registered MRI volumes, we propose to constrain the random initial position (x', y', z') to be within a fixed search window centered around the current voxel position (x, y, z). Then, for each voxel in S, an index template t is assigned using *i.i.d.* random variable within $\{1, \ldots, n\}$. Consequently, each patch in S is associated to a unique random match among all templates of the library T. Considering the important number of patches in S, all templates are very likely to be reached at least once. Moreover, although the corresponding template is randomly selected during the initialization step, all matches can move from a template to another during the following iterative process. Figure 1(a) illustrates the initialization step. For each patch in S (only three are displayed), the fixed search window for the random initialization is depicted in dotted lines in the different training templates.

This constraint has two advantages. First, it improves the matching convergence, making good use of the linear registration between training template and the subject. Second, limiting the initialization to a fixed window prevents



(a) CI (b) PS for iteration #1 (c) CRS for iteration #1

(d) PS for iteration #2 (e) CRS for iteration #2 (f) multiple OPM

Figure 1: Optimized PatchMatch (OPM) main steps. In this figure, the representation of OPM steps focuses on the blue patch in S. Green, pink, purple and orange colors represent the adjacent patches of the blue patch. During the constrained initialization (CI) (a), patches of the subject S are matched (full lines) to a random patch of the library within an initialization search window (three are displayed). The propagation step (PS), is represented for iteration #1 and #2 in (b) and (d), respectively. The shifted correspondences of recently processed adjacent patches are tested for improvement (dotted lines). Constrained random search (CRS) for iteration #1 and #2 are represented for the blue patch, in (c) and (e), respectively. Random tests are performed within a decaying search window around the current best match, within the current best template. In (f), the result of multiple independent ANN searches by OPM is illustrated. See text for more details.

the algorithm from finding similar patches in terms of intensity (low SSD) that are spatially far, leading to potential segmentation errors. As a consequence, our constrain initialization reinforces spatial proximity between voxels in S and their matches in T and makes the algorithm converge faster.

As in the original PatchMatch algorithm, after this constrained initialization, propagation and random search steps are performed iteratively in order to improve the patch correspondence.

2.1.4. Propagation Step with Fast Distance Computation

The propagation step of OPM is the 3D extension of the one proposed in [22]. For each patch located at $(x, y, z) \in S$, an ANN improvement is performed by testing if the shifted ANN of its 6 directly adjacent patches located at $(x \pm 1, y, z)$, $(x, y \pm 1, z)$ and $(x, y, z \pm 1)$ provides a better match. In order to converge faster and to propagate good correspondences, the original PM only tests recently processed neighbors during this step. Consequently, in 3D, only three adjacent neighbors are tested at each iteration, according to the raw scan order. Figures 1(b) and 1(d) illustrate this step, where the blue dotted lines correspond to the test of shifted adjacent neighbors in T, in order to improve the current blue patch correspondence. In this example, the best match for the blue patch moves from template T_1 to T_2 with iteration #1 and from T_2 to T_1 with iteration #2. The propagation step is a core stage of the OPAL algorithm since it allows a patch correspondence to move over all the templates in T. Thus, the ANN of the current voxel can move from one template to another one, since the ANN of the adjacent voxels are not necessarily in the same template.

Moreover, the computational burden of these tests can be extremely reduced in the propagation step. Indeed, we propose an acceleration technique based on the observation that the ANN of the adjacent patches are known. As neighbor patches are overlapping, we use a shifted SSD instead of computing the whole distance between the current patch and the shifted ANN of its adjacent patch. Hence, only the non overlapping coordinates are considered, *i*.e., the two squares at 3D patches extremities, since there is a one voxel shift in only one of the

three dimensions. The exact SSD between the current patch and the shifted correspondence is thus obtained in the fastest way. The patch overlapping is illustrated in Figure 1(b), where the blue square overlaps the green and pink ones. The distances on the overlapping areas do not need to be re-computed.

2.1.5. Constrained Random Search

In the original PM algorithm [22], the random search step is performed on all dimensions. In contrast to the original method, OPAL deals with a library of images. Therefore, we modify the random search step to take into account this aspect. In order to ensure spatial consistency, OPAL performs the random search only in the current template containing the current best patch correspondence (*i.e.*, t is fixed, and we random on $(x'_t, y'_t, z'_t) \in T_t$) within a search window decaying by a factor 2. The process stops when the window is reduced to a single voxel. The decaying search window size is empirically defined as the size of the initialization window. Figures 1(c) presents examples of such fixed template random search where the decaying search windows are represented in dotted blue lines.

2.1.6. Multiple PM and Parallel Computation

Contrary to [22] that only estimates the best match with PM, OPAL computes k-ANN matches in T. These ANNs are then used to perform the label fusion. In the literature, an extension of the original PM algorithm to k-ANN case has been proposed in [24]. The suggested strategy is to build a stack of the best visited matches. At each new tested match, the distance is compared to the one of the worst ANN among the stack. If there is an improvement in terms of SSD, the worst ANN is replaced by the new match. However, to parallelize such an approach, the current image S must be split into several parts. Since PM uses propagation of good matches between adjacent patches, any split would lead to boundary issues. Therefore, in OPAL, we decide to implement the k-ANN search through k independent OPM, denoted as k-OPM. This leads to a more efficient and simple multi-threading. Consequently, each thread can run

an OPM without any dependencies to the other ones. Figure 1(f) illustrates the result of the multiple OPM steps with $k = 3$. One can note that k independent OPM can lead to the same ANN for a given voxel. The redundancy of the same ANN in the ANN map is not an issue, since each contribution is weighted during the patch-based label fusion step. During our validation, for the considered size of training libraries, we experimentally observed that such multiple selections of the same ANN is a rare phenomena.

2.2. Patch-based Segmentation

After convergence of the multiple OPM, the position and the distance of the k-ANN is known. Therefore, a patch-based label fusion step can be used to produce the final segmentation. In such a method, labels are fused according to their relevance to compute an estimator map of the subject to segment. In contrast to the original PBL method [15], where only the central voxel information was considered, OPAL segmentation is performed in a patchwise manner, using the whole training patch as done in [16, 21, 25]. Moreover, as recently proposed in [25], OPAL uses a bilateral kernel for weight computation in order to reinforce spatial coherency. Figure 2 illustrates the patch-based label fusion process and the computation of the estimator map and is detailed below.

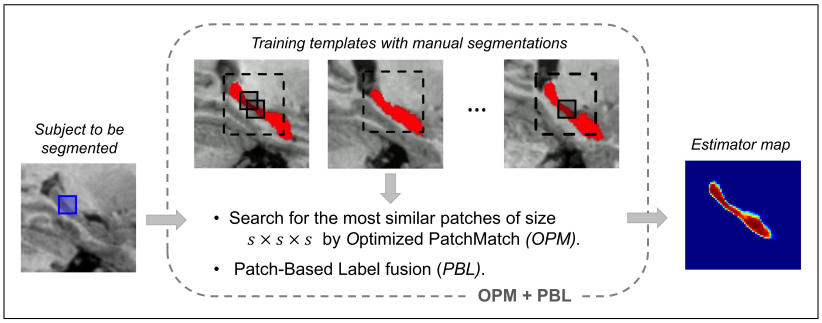


Figure 2: Core of OPAL method: optimized PatchMatch and patch-based label fusion on image intensities. For every voxel of the subject to segment, a search for similar patches of size $s \times s \times s$ is carried out by OPM. A patch-based label fusion is then performed to generate a label estimator map. See text for more details.

2.2.1. Patchwise Label Fusion

At the end of the matching process, the k-ANN are estimated for all the patches in S. Thus, the location and the SSD between the patches of S and their k-ANN in T are known. To obtain the final segmentation, we use the Patch-based label fusion (PBL) method presented in [15]. In contrast to [15], that considers all the patches within a fixed number of preselected templates, OPAL only uses the k most similar patches (limiting segmentation error) over the entire library (increasing segmentation accuracy). As previously mentioned, when the same ANN is selected several times by independent PM, it will be taken into account several times during the label fusion. Considering a 3D patch $\mathcal{P}(\mathbf{x_i})$ at voxel position $\mathbf{x_i} = (x, y, z) \in S$, and $\mathcal{K}_i = \{\mathbf{x_{j,t}}\}$ the set of its k-ANN match positions, its label fusion $\mathcal{L}(\mathbf{x_i})$ is defined by,

$$\mathcal{L}(\mathbf{x_i}) = \frac{\sum_{\mathbf{x_{j,t}} \in \mathcal{K}_i} \omega(\mathbf{x_i}, \mathbf{x_{j,t}}) l(\mathbf{x_{j,t}})}{\sum_{\mathbf{x_{j,t}} \in \mathcal{K}_i} \omega(\mathbf{x_i}, \mathbf{x_{j,t}})}, \tag{2}$$

where $\omega(\mathbf{x_i}, \mathbf{x_{j,t}})$ is the weight assigned to $l(\mathbf{x_{j,t}})$, the binary label given by the expert at voxel $\mathbf{x_{j,t}} = \{\mathbf{x_j}, t\} \in T$.

The weight $\omega(\mathbf{x_i}, \mathbf{x_{j,t}})$ depends on the similarity between the patches $\mathcal{P}(\mathbf{x_i}) \in S$, the patch contributing to the labeling of $\mathbf{x_i}$, and the ANN patch $\mathcal{P}(\mathbf{x_{j,t}}) \in T$. This weight is defined as,

$$\omega(\mathbf{x_i}, \mathbf{x_{j,t}}) = \exp\left(1 - \frac{\|\mathcal{P}(\mathbf{x_i}) - \mathcal{P}(\mathbf{x_{j,t}})\|_2^2}{h(\mathbf{x_i})^2}\right), \tag{3}$$

where $h(\mathbf{x_i})^2 = \alpha^2 \min_{\mathbf{x_{j,t}} \in \mathcal{K}_i} (\|\mathcal{P}(\mathbf{x_i}) - \mathcal{P}(\mathbf{x_{j,t}})\|_2^2 + \epsilon)$, with ϵ a small constant to ensure numerical stability, and α a normalization constant. With the parameter $h(\mathbf{x_i})$, the distance of the current contribution is divided by the minimal distance among all k-ANN contributions.

Most nonlocal label fusion methods perform voxelwise aggregation, which can provide a lack of regularization on final segmentation. Therefore, to further improve segmentation quality, the label fusion is performed over the whole patch as done in [16, 21, 25] and not only using the central voxel. The patchwise

labeling is then computed as follows,

$$\mathcal{L}(\mathcal{P}(\mathbf{x_i})) = \frac{\sum_{\mathbf{x_{j,t}} \in \mathcal{K}_i} \omega(\mathbf{x_i}, \mathbf{x_{j,t}}) l(\mathcal{P}(\mathbf{x_{j,t}}))}{\sum_{\mathbf{x_{j,t}} \in \mathcal{K}_i} \omega(\mathbf{x_i}, \mathbf{x_{j,t}})}. \quad (4)$$

This way, 3D patches $\mathcal{P}(\mathbf{x_i}) \in S$ are labeled at the same time. At the end, the label estimator for voxel $\mathbf{x_i}$ is obtained by averaging all neighbors contributions from overlapping blocks containing $\mathbf{x_i}$ to obtain the estimator map $\mathcal{F}$.

2.2.2. Bilateral Kernel

In addition to the patchwise strategy, a spatial filtering is performed during segmentation in order to reinforce spatial coherency of the selected k-ANN. The spatial filtering exploits the observation that structures of interest are spatially close due to the linear registration. Therefore, good patch candidates should be similar in term of intensity and spatially not too far. Therefore, as done in NICE [25], each ANN contribution to patchwise labeling is also weighted by the spatial distance between patch centers $\mathbf{x_i} \in S$ and $\mathbf{x_{j,t}} = \{\mathbf{x_j}, t\} \in T$,

$$\omega(\mathbf{x_i}, \mathbf{x_{j,t}}) = \exp\left(1 - \left(\frac{\|\mathcal{P}(\mathbf{x_i}) - \mathcal{P}(\mathbf{x_{j,t}})\|_2^2}{h(\mathbf{x_i})^2} + \frac{\|\mathbf{x_i} - \mathbf{x_j}\|_2}{\sigma^2}\right)\right), \quad (5)$$

where σ^2 is a normalization constant.

2.3. Late Aggregation of Multi-Scale and Multi-Feature Estimators

Due to the high computational cost of previously published multi-templates methods, most were designed in a mono-scale and mono-feature context. Recently, multi-scale [19, 26, 27], and multi-feature [28, 29] approaches have been investigated. These studies show the advantage of such frameworks. However, since these methods require a non negligible computation time, they are based on either multi-scale [19, 26, 27] or multi-feature [28, 29] estimation but not both at the same time. Moreover, these methods perform early feature aggregation: all the considered scales or features are fused into a single vector before performing patch comparison. However, early fusion is not necessarily the best strategy. Usually used for computation time consideration, early fusion has been shown to be less efficient than late estimator fusion/aggregation [30]. Moreover, the

use of both multi-scale and multi-feature should improve segmentation accuracy. Leveraging the computational efficiency of OPAL, we propose to investigate a new framework to simultaneously perform multi-scale and multi-feature analysis with late aggregation of estimators. Figure 3 illustrates the whole OPAL method and the late fusion of multi-feature and multi-scale label estimator maps.

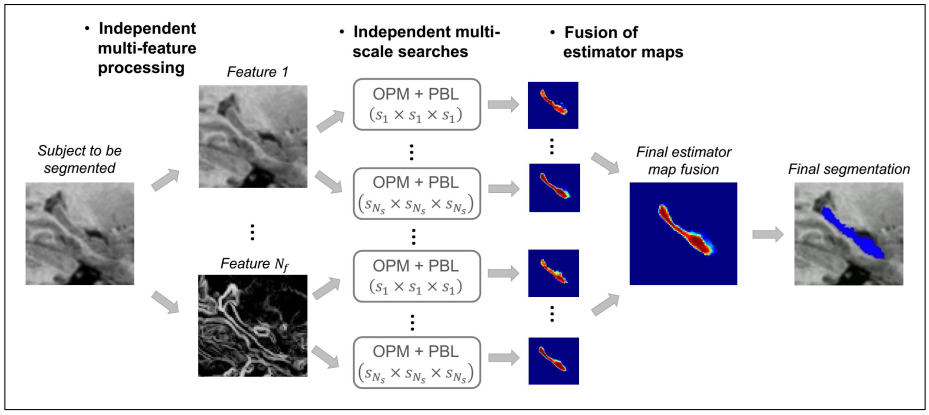


Figure 3: OPAL method. Fusion of multi-feature and multi-scale label estimator maps. The algorithm is applied with N_s different patch sizes, on N_f different features, so $N = N_s \times N_f$ estimator maps are computed and merged to provide the final segmentation. See text for more details.

2.3.1. Multi-scale Estimators

In patch-based methods, the structure description highly depends on the size of the patch. The patch size needs to be large enough to capture the local geometry and to prevent discontinuities in the segmentation. However, using very large neighborhoods may reduce the probability of finding similar patches in the library. Although the optimal patch size can be determined by experiments for a given dataset, multi-scale approaches may significantly improve segmentation accuracy as shown in recent multi-scale label fusion approaches [26, 27]. In these papers, the ANN search consists in finding the candidate minimizing the distance for every scale at the same time. Therefore, such a strategy selects a consensual candidate providing the best similarity on average over all the considered scales. In contrast to these previous works, we propose to perform

fully independent multi-scale ANN searches where a candidate providing the best similarity is obtained for each scale. With this method, k-OPM are independently computed for multiple patch sizes $s_i, i \in \{1, \ldots, N_s\}$. Consequently, in our context, multi-scale refers to the simultaneous use of patches of different sizes, and the images are considered with their initial resolution. In Figure 3, the ANN search by OPM and PBL are performed on each feature for N_s patch sizes.

2.3.2. Multi-feature Estimators

Similarly, the search for similar patches by OPM can also be carried out independently on different features (edges, textures, etc.). During our tests with different potential features, we found that using the gradient norm (*i.e.*, first intensity derivative) in addition to the original MRI intensities increases the segmentation accuracy. Therefore, we use both these features. Figure 3 shows how OPAL is applied to the N_f features extracted from the subject S to segment. The resulting estimator maps are then merged *a posteriori* as explained in the next section. As for the multi-scale aspect, our framework contrasts with recent multi-feature methods [29] where the ANN search consists in finding the best candidate for every feature at the same time. In our method, the independent searches improve the ANN diversity of the selected matches.

2.3.3. Late Aggregation of Estimators

Label estimator maps are independently computed from PBL on multi-scale and multi-feature ANN searches. The last step is the aggregation of these estimator maps to generate the final segmentation. Here, OPAL is applied on N_f features, with N_s different patch sizes, so $N = N_s \times N_f$ estimator maps $\mathcal{F}^i$ with $i \in \{1, \ldots, N\}$ are computed to generate the final segmentation. The final estimator map $\mathcal{F}$ is then computed by averaging the estimator maps by a late fusion [30],

$$\mathcal{F} = \frac{\sum_{i=1}^{N} \mathcal{F}^i}{N}. \tag{6}$$

In the end, the final label decision is taken as follows:

$$\mathcal{M}(\mathbf{x_i}) = \begin{cases} 1, & \text{if } \mathcal{F}(\mathbf{x_i}) \geq 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

3. Materials

3.1. Datasets

During our experiments on hippocampus segmentation, two different datasets have been considered. We used images from elderly adults obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [31] and images from young adults obtained from the International Consortium for Brain Mapping (ICBM) dataset [32]. Our goal was to demonstrate the robustness of our OPAL framework using data from different sources with different preprocessing pipelines.

EADC-ADNI. This dataset was used to evaluate the performance of our approach. The European Alzheimer's Disease Consortium and Alzheimer's Disease Neuroimaging Initiative (ADNI) Harmonized Protocol (HarP) is a Delphi definition of manual hippocampus segmentation from MRI that can be used to validate automated segmentation algorithms [33]. The EADC-ADNI dataset is based on ADNI MRI scans [31] which were acquired on General Electric, Philips, and Siemens scanners using a 3D MPRAGE T1-w sequence as recommended by the MRI Core of the ADNI consortium. The ADNI acquisition protocol is based on sagittal 3D MP-RAGE sequence (TR=2400ms, minimum full TE, TI=1000ms, FOV=240mm, voxel size of $1.25\times1.25\times1.2\text{mm}^3$). Images were then reconstructed at a voxel size of approximately $1\times1\times1.2\text{mm}^3$. As part of the EADC-ADNI, 100 MRI of the ADNI dataset have been manually labeled according to the harmonized protocol and are freely available (`www.hippocampal-protocol.net`). The definition of the harmonized protocol has been designed to reduce inconsistencies of manual segmentation protocols as detailed in [33]. The mean Dice value for repeated manual segmentations between experts has been estimated to 89% ([88%; 92%]) accord-

ing to [34]. All the images were preprocessed using the volBrain pipeline (`http://volbrain.upv.es`). The first preprocessing step is based on the adaptive nonlocal mean filter [35]. Denoised MRI are then coarsely corrected for inhomogeneity with N4 [36]. Afterwards, an affine registration to MNI space is achieved using ANTS [37]. In the MNI space, a fine inhomogeneity correction is performed using SPM8 routines [38]. Finally, an intensity normalization procedure is applied to the images [39]. The whole preprocessing pipeline is performed in less than 5min per subject.

ICBM. We used a part of the International Consortium for Brain Mapping (ICBM) dataset [32] which consists of 80 MR images of young and healthy individuals with manual segmentations following the Pruessner's protocol [40]. The MRI scans were acquired with a 1.5T Philips GyroScan imaging system (1mm thick slices, TR=17ms, TE=10ms, flip angle=30 °, FOV=256mm). The estimated intra-class reliability coefficient was of 90% for inter- (4 raters) and 92% for intra-rater (5 repeats) reliability. All the images were preprocessed through the following pipeline: estimation of the standard deviation of noise [41]; denoising using the optimized nonlocal means filter [42]; correction of inhomogeneities using N3 [43]; registration to stereotaxic space based on a linear transform to the ICBM152 template ($1\times1\times1\text{mm}^3$ voxel size) [44]; linear intensity normalization of each subject on template intensity; image cropping around the structures of interest; and cross-normalization of the MRI intensity between the subjects with [39]. As for EADC-ADNI preprocessing, the whole pipeline requires less than 5min per subject.

3.2. Quality Metric and Compared Methods

The proposed method was validated through a leave-one-out cross validation procedure for both datasets. The segmentation accuracy was estimated with the standard Dice coefficient (also called kappa index) introduced in [45] which compares the expert-based segmentation with the automatic segmentation. For

two binary segmentations $\mathcal{M}_1$ and $\mathcal{M}_2$, the Dice coefficient D is computed as,

$$D(\mathcal{M}_1, \mathcal{M}_2) = \frac{2 \mid \mathcal{M}_1 \cap \mathcal{M}_2 \mid}{\mid \mathcal{M}_1 \mid + \mid \mathcal{M}_2 \mid}. \quad (8)$$

For each subject, the Dice coefficient of left and right hippocampus are averaged and the values in Tables 1, 2 and 3 correspond to the median Dice over all the dataset. The associated computation times include ANN map computation for every feature with every patch size, PBL on every estimator map and final segmentation of both left and right hippocampus. During our validation process, we investigated the impact of parameters such as the initialization search window size, the patch size, the number of neighbors (*i.e.*, number of OPM), and the impact of multi-scale and multi-feature approaches on segmentation accuracy and computation time.

The results obtained by OPAL were compared to the published results on the ICBM dataset of the original Patch-Based Label fusion method (PBL) [15], a Sparse Representation Classification method (SRC) [20], and a dictionary learning method, denoted as Discriminative Dictionary Learning for Segmentation (DDLS) [20]. Mean Dice coefficients of left and right hippocampus results of EADC-ADNI dataset were compared to the results obtained with a Random Forest approach [34], and two multi-templates based approaches, BioClinica Multi-Atlas Segmentation algorithm (BMAS) [46], and Learning Embeddings for Atlas Propagation (LEAP) [47].

3.3. Implementation Details

OPAL was implemented in MATLAB using multi-threaded C-MEX code. Our experiments were carried out using a server of 16 cores at 2.6 GHz with 100 GB of RAM. Default parameters are set to process both ICBM and ADNI datasets. These parameters offer a good trade-off between segmentation accuracy and computation time. In the following results, OPAL is processed with 3 inner iterations of OPM and the number of threads on each feature is equal to k. In (5), parameters α and σ are empirically set to 2. In the multi-feature setting,

estimator maps are computed from image intensities and gradient norm intensities. In the multi-scale setting, OPAL is processed with 3×3×3 and 5×5×5 voxels patch sizes on each feature. Finally, the number of selected matches per voxel for each estimator is by default set to $k = 10$ ANNs, and the size of the initialization search window is set to 13×13×13 voxels.

4. Results

4.1. Influence of Parameters

First, as mentioned in 2.1.3, the initialization search window reinforces spatial coherency between voxels in S and their matches in T. By setting the optimal search window area, the algorithm converges faster since more relevant matches are found, thus leading to a higher segmentation accuracy. This optimal window size is empirically estimated according to the dataset. Figure 4 shows the Dice coefficient for several initialization window sizes on both studied datasets. For ICBM, a plateau is reached for a search window of 7×7×7 voxels, while an area of 13×13×13 voxels leads to better segmentation results for the EADC-ADNI dataset. This second dataset requires a larger search window size since it contains higher anatomical variability due to the presence of pathologies. Therefore, in the following, the initialization window is by default set to 13×13×13 voxels.

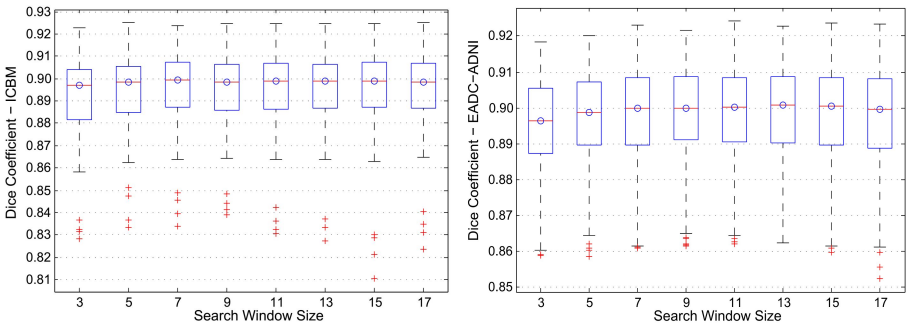


Figure 4: Influence of the initialization search window on Dice coefficient for the ICBM (left) and the EADC-ADNI (right) datasets.

Figures 5 and 6 show the influence of the number of ANN (*i.e.*, k) and of the patch size on the segmentation quality and on the computation time. Without the multi-scale approach, we found out that patches of size 5×5×5 voxels provide the best results on both datasets. This patch size indeed gives acceptable description for structures of different scales, as already observed in [15, 20]. With our multi-scale approach, we can automatically take advantage of different patch sizes that provide better results. By merging estimator maps generated from 3×3×3 and 5×5×5 voxels patch sizes, we reach a Dice coefficient of 89.9% for the ICBM dataset, with default settings. (*i.e.*, k=10 ANNs, multi-scale, multi-feature and initialization window set to 13×13×13 voxels). By adding estimator maps from 7×7×7 voxels patch sizes and increasing the number of k-OPM, we even reach a 90.1% Dice coefficient. For the EADC-ADNI dataset, we reach a 90.1% Dice coefficient (90.05% with default parameters). For both datasets, the segmentation step is performed in less than 2s of processing per subject. These results highlight the importance of taking into account the diversity of information obtained from various patch sizes. We noted that the median Dice coefficient reaches a plateau around 10-ANN. It is interesting to note that this number is coherent with the suggested number of templates in multi-template matching methods [7]. As expected, bigger patches and larger number of ANN require higher computation time. Consequently, our experiments suggest that using $k = 10$ ANNs on each feature offers a good trade-off between segmentation accuracy and computation time.

Different settings were compared using paired t-test on Dice coefficients. The results in Tables 1 and 2 present the impact of each contribution on Dice coefficient and computation time during the segmentation process. For both datasets, the use of multi-feature and multi-scale significantly improved the segmentation accuracy compared to mono-scale and mono-feature method, as assessed by p-values. Moreover, in all studied cases, multi-scale and multi-feature approaches improved results of mono-scale and multi-feature method. This demonstrates the complementary nature of the multi-feature and multi-scale strategy.

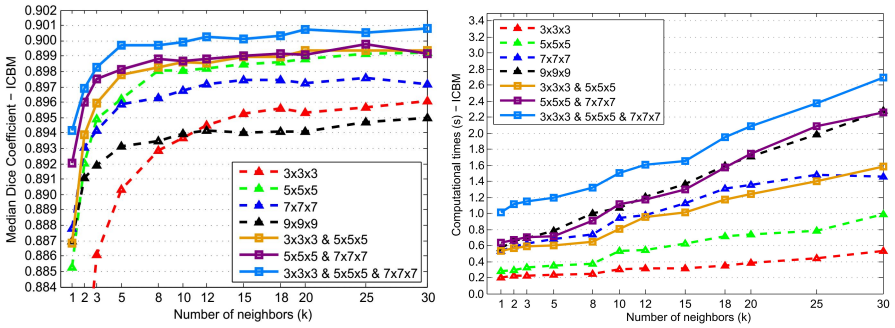


Figure 5: Median Dice coefficient according to the mono-scale and multi-scale patch sizes and the number of neighbors (left), and the corresponding computation time (right) for the ICBM dataset. These results are obtained with default multi-feature settings, *i.e.*, MRI gradient norm in addition to the original MRI intensities.

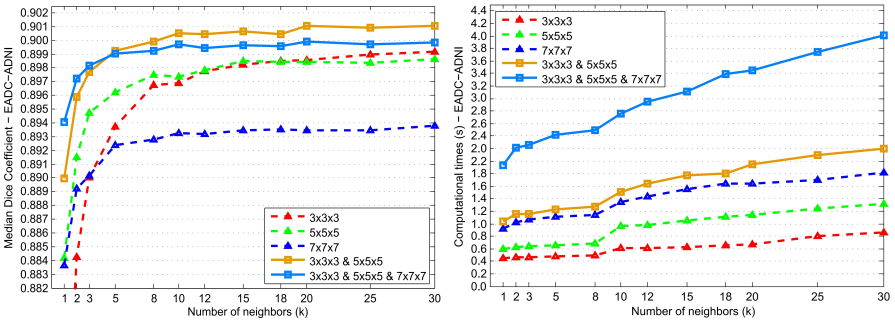


Figure 6: Median Dice coefficient according to the mono-scale and multi-scale patch sizes and the number of neighbors (left), and the corresponding computation time (right) for the EADC-ADNI dataset. These results are obtained with default multi-feature settings, *i.e.*, MRI gradient norm in addition to the original MRI intensities.

Estimator maps for several features and several patch sizes are shown in Figure 7, for a subject of the EADC-ADNI dataset. First, bigger patch sizes produce smoother estimator maps. Smaller patches are able to better capture finer details at the expense of noisier estimator maps. Second, the estimators based on gradient norm better define edge structure but are less robust to noise. Finally, the aggregation is able to produce a good trade-off between considered scales and features.

Figure 8 presents segmentation results of best, median and worst subjects

obtained on the EADC-ADNI dataset. First, we can see that automatic method produces a smoother segmentation than expert. The patchwise label fusion obtains consistent segmentation along the edge, but tends to fill holes present in manual segmentation. Some of these holes appear to be hippocampal CSF while others seem to be expert inaccuracies.

OPAL on ICBM	Median Dice	Mean Dice	p-value	Comp. Time
Mono-scale, Mono-feature	89.4%	$89.1 \pm 1.85\%$	$< 10^{-14}$	$0.27s$
+ Multi-feature	89.8%	$89.6 \pm 1.68\%$	0.0131	$0.53s$
+ Multi-scale	89.9%	$89.7 \pm 1.70\%$	$\times$	$0.92s$

Table 1: Influence of multi-scale and multi-feature in terms of segmentation accuracy and computation time on the ICBM dataset. Mono-scale and mono-feature results are obtained with PBL from 5×5×5 voxels patch size ANN search on MRI intensities. Multi-feature considers the MRI gradient norm in addition to the original MRI intensities. Multi-scale adds estimator maps computed from 3×3×3 voxels patch sizes on each feature. The given computation times correspond to the mean segmentation processing time of one subject.

OPAL on EADC-ADNI	Median Dice	Mean Dice	p-value	Comp. Time
Mono-scale, Mono-feature	89.4%	$89.2 \pm 1.55\%$	$< 10^{-25}$	$0.49s$
+ Multi-feature	89.7%	$89.6 \pm 1.45\%$	$< 10^{-8}$	$0.95s$
+ Multi-scale	90.1%	$89.8 \pm 1.46\%$	$\times$	$1.51s$

Table 2: Influence of multi-scale and multi-feature in terms of segmentation accuracy and computation time on EADC-ADNI dataset. Mono-scale and mono-feature results are obtained with PBL from 5×5×5 voxels patch size ANN search on MRI intensities. Multi-feature considers the MRI gradient norm in addition to the original MRI intensities. Multi-scale adds estimator maps computed from 3×3×3 voxels patch size on each feature. The given computation times correspond to the mean segmentation processing time of one subject.

4.2. Comparison with State-of-the-Art Methods

The performances obtained by OPAL are compared to other methods applied to the same dataset in Tables 3 and 4. The presented values are the results published by the authors. The provided computation times are the times dedicated to segmentation step only but do not include template preselection while

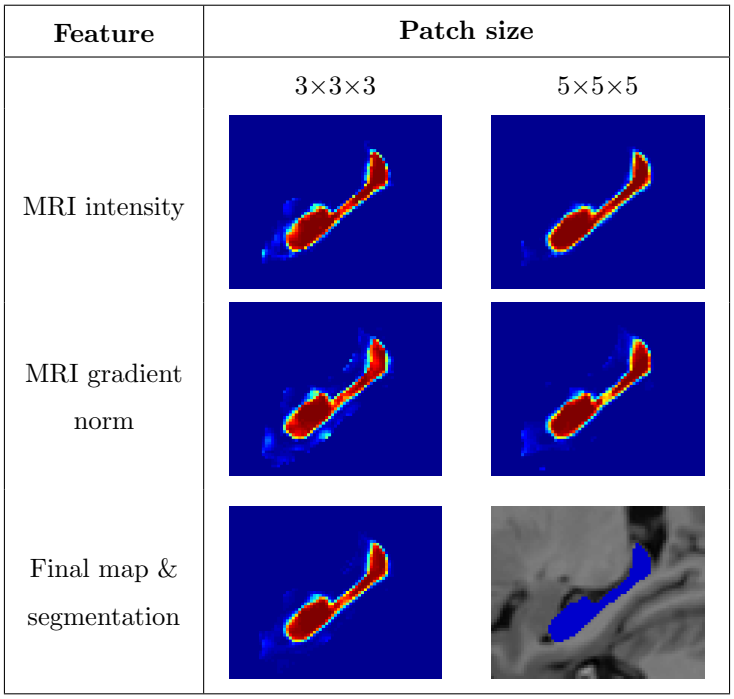


Figure 7: 2D visualizations of estimator maps for several features and several patch sizes for the EADC-ADNI dataset. With patches of size 5×5×5 voxels, estimator map decision is more stable for every voxel (higher intensity within the hippocampus volume). With patches of size 3×3×3 voxels, some areas are more accurately segmented, see for instance the peak on top on the hippocampus image.

only OPAL does not require it. Therefore, the computation times are underestimated except for OPAL.

On the ICBM dataset, compared to the original PBL [15], OPAL improves segmentation accuracy by 1.7 percentage points (pp) while being 700× faster. Compared to the most accurate method on this dataset, based on dictionary learning (DDLS [20]), OPAL obtained higher Dice coefficients for computation times 1000× faster and with a p-value inferior to 10^{-12} obtained from a paired t-test on the OPAL and DDLS sets of Dice coefficients. In addition, for a given Dice coefficient of 89.0% (equivalent to the DDLS method accuracy) OPAL

requires less than 0.22s on the ICBM dataset (4000× faster than DDLS method).

On the EADC-ADNI dataset, OPAL results are compared to other methods only in terms of segmentation accuracy, since computation times are not provided by the authors in their publications. The results presented with OPAL on EADC-ADNI in Table 4 are obtained in 1.51s processing per subject. In all studied cases, OPAL produced the best segmentation accuracy with a mean Dice coefficient of 89.8% (median Dice of 90.1%). The Dice values show that OPAL outperforms recently proposed methods on EADC-ADNI. Indeed, compared to a Random forest approach [34], OPAL improves segmentation accuracy by 13.8pp and compared to recent multi-template approaches OPAL obtained a gain superior to 2.2pp, with a p-value inferior to 10^{-25} obtained from a paired t-test on the OPAL and LEAP sets of Dice coefficients.

Method on ICBM	Median Dice	95% interval	Comp. Time
Patch-based (PBL)[15]	$88.2 \pm 2.19\%$	$[87.7; 88.7]\%$	$662s$ ($\times 700$)
Multi-templates (MTM)[7]	$88.6 \pm 2.05\%$	$[88.2; 89.0]\%$	$3974s$ ($\times 4300$)
Sparse coding (SRC)[20]	$88.7 \pm 1.94\%$	$[88.3; 89.2]\%$	$5587s$ ($\times 6000$)
Dictionary learning (DDLS)[20]	$89.0 \pm 1.90\%$	$[88.5; 89.4]\%$	$943s$ ($\times 1000$)
OPAL	$\mathbf{89.9 \pm 1.70}\%$	$[\mathbf{89.6}; \mathbf{90.3}]\%$	**0.92s**

Table 3: Method comparison in terms of segmentation accuracy and computation time (per subject) for the ICBM dataset.

Method on EADC-ADNI	Mean Dice	95% interval
Random Forest [34]	$76.0 \pm 7.00\%$	$[74.6; 77.4]\%$
Multi-templates (BMAS)[46]	$86.6 \pm 1.70\%$	$[86.3; 86.9]\%$
Multi-templates (LEAP)[47]	$87.6 \pm 2.07\%$	$[87.1; 88.0]\%$
OPAL	$\mathbf{89.8 \pm 1.46}\%$	$[\mathbf{89.5}; \mathbf{90.1}]\%$

Table 4: Method comparison in terms of segmentation accuracy for the EADC-ADNI dataset. Since none of the selected publications mention their computation times, the comparison only focus on the mean Dice coefficient. The selected result for OPAL method was obtained in 1.51s processing per subject.

Automatic segmentations as priors. Recently, several works have proposed to use automatic segmentations as priors in order to accurately segment a new subject. A way to improve segmentation accuracy consists in increasing the size of the template library. In order to do this, subjects without expert segmentations are automatically segmented and added to the template library of manually segmented subjects [19]. The Multiple Automatically Generated Templates (MAGeT) approach has been proposed in [48] and works by propagating segmentations to a template library, composed of a subset of unlabeled subjects, via transformations estimated by nonlinear registrations. The resulting segmentations are then used as template library to segment a new subject. Similarly, the LEAP method [47] proposes to propagate the label segmentation to unlabeled subjects by iteratively segmenting the closest subjects in terms of joint entropy. These approaches lead to segmentation accuracy improvement, since the diversity of the dataset used to segment a subject is increased.

As mentioned in section 2.1.2, the computation time and complexity of OPAL only depends on the size of the subject to segment. This important fact enables us to extend the library size with no impact on the complexity of the algorithm. New subjects without manual expert segmentations can be automatically segmented and added to the template library in order to improve its diversity. Consequently, the segmentation accuracy of a new subject may be improved, since more relevant matches can be found within the template library.

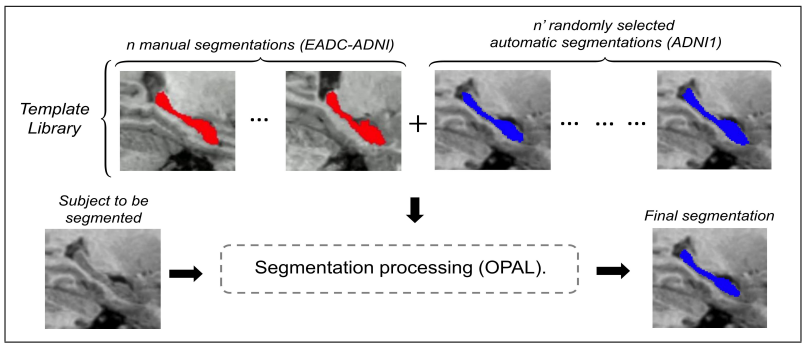


Figure 9: Addition of new segmented subjects to the template library. The automatic segmentation of new subjects provided without manual expert segmentations can be added to the template library in order to increase its size and diversity. Consequently, later segmentations may benefit from more numerous and potentially better training templates.

Therefore, we propose an experiment where automatically segmented subjects from the standardized ADNI1 dataset [49] are randomly selected and added to the EADC-ADNI template library as illustrated in Figure 9. The Dice coefficient is still computed with a leave-one-out procedure on the EADC-ADNI subjects with provided expert-based segmentations. Figure 10 shows the impact of increasing the library size, on the segmentation accuracy and computation time.

Adding new templates to the library with automatic segmentations as priors enables us to improve the segmentation accuracy. Indeed, since the dataset is extended with new subjects, its diversity is increased and more relevant matches can be found within the template library. Most importantly, the computation time results in Figure 9 highlight the important fact that OPAL complexity only depends on the size of the subject to segment and not on the size of the template library. Adding subjects to the database improves the segmentation accuracy at the expense of a very little setback on computation time (due to memory storage and data transfer). With 50% of supplementary training templates, the computation time is only increased by 6%.

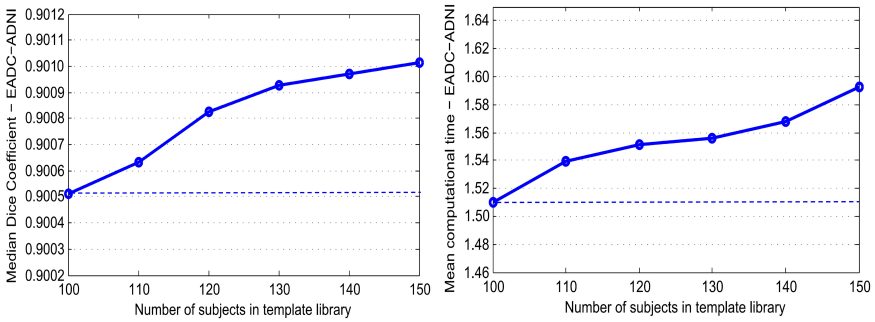


Figure 10: Influence of the addition of automatic segmented ADNI subjects to the EADC-ADNI dataset on the segmentation accuracy (left) and the corresponding computation time (right). The results obtained with 100 subjects (dotted line) correspond to the selected results in Table 2.

Clinical application. Finally, we propose to show the performance of our method on a clinical application, by comparing population separation accuracy using manual segmentation of the EADC-ADNI harmonized protocol (HarP) [33] and the OPAL segmentation. The area under the ROC curve (AUC) is computed on hippocampal volumes in the MNI space for both manual and OPAL segmentation results on the three groups of the EADC-ADNI dataset, AD (Alzheimer's Disease, N=37), MCI (Mild Cognitive Impairment, N=34) and NC (Normal Controls, N=29). As shown in Table 5, the segmentation results provided by OPAL enable to better separate groups with a higher AUC. The Pearson's correlation is also computed between the HarP and OPAL hippocampal volumes of segmentations. In Figure 11, the hippocampal volumes distribution for each group are represented. The correlation between hippocampal volumes of HarP and OPAL segmentations is also illustrated.

	EADC-ADNI HarP	OPAL
HC mean volume (mm^3)	9397 ± 1588	9272 ± 1525
AUC NC vs. AD	0.884	0.898
AUC NC vs. MCI	0.805	0.821
AUC MCI vs. AD	0.612	0.634

Table 5: Area under the ROC curve (AUC) on hippocampal volumes in the MNI space of the segmentation results from reference EADC-ADNI harmonized protocol and OPAL method.

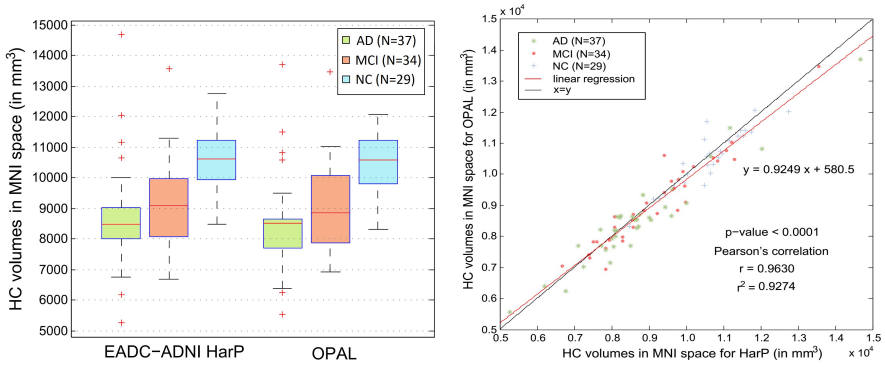


Figure 11: Hippocampal volumes in the MNI space of the segmentation results from reference EADC-ADNI harmonized protocol and OPAL method (left). Correlation between hippocampal volumes of HarP and OPAL segmentations (right).

5. Discussion

Our proposed OPAL method presents several differences with state-of-the-art PBL approaches. First, the complexity of the optimized PatchMatch algorithm (see Figure 1) only depends on the size of subject's image. Consequently, the entire image library T is used without any template preselection step, at constant complexity in time. The linear registration is also exploited by constraining the search for patch matches at each step. Secondly, a patchwise label fusion is performed from the selected matches (see Figure 2) and a bilateral kernel is also used to increase spatial consistency leading to better segmentation results, as done in [25]. Finally, we introduced a new multi-scale and multi-feature

framework based on late aggregation of estimators. This new approach is possible thanks to the very low computational burden of the ANN search in our OPM framework. Independent multi-scale and multi-feature ANN searches are carried out, and a late fusion is finally performed on all resulting estimator maps from PBL to produce the final segmentation as illustrated in Figure 3. We validated our method on two datasets for hippocampus segmentation. These datasets cover different manual segmentation protocols and preprocessing pipeline. By this way, the robustness of OPAL to hippocampus definition and processing has been studied.

On ICBM and EADC-ADNI datasets, we respectively obtained a median Dice coefficient of 89.9% and 90.1% in approximately 1.5s processing per subject. A large comparison with published methods such as original PBL [15], sparse representation (SRC) [20], dictionary learning (DDLS) [20], multi-templates (MTM, BMAS, LEAP) [7, 46, 47] and random forest [34], highlights the very competitive results of the proposed method (see Tables 3 and 4).

For the EADC-ADNI comparison, the computation times are not provided by the authors. However, we may assume that the BMAS [46] and LEAP [47] methods are likely to propose comparable computation time to MTM [7] since they are also based on a multi-templates warping approach. One can note that multi-templates warping methods perform worse on the EADC-ADNI dataset than on the ICBM dataset. This can be related to higher anatomical variability in EADC-ADNI dataset due to the presence of Alzheimer's disease (AD). On this dataset, the well defined one-to-many mapping offering by patch-based segmentation appears to better capture this higher variability.

It is important to note OPAL can reach the inter-expert reliability on both datasets (90% and 89.0% respectively for ICBM and EADC-ADNI datasets). Moreover, this has been validated on two datasets with two different manual segmentation protocols. While more than 30 minutes are required by an expert to segment one hippocampus (1 hour for both), OPAL produces similar segmentation quality in less than 2s. OPAL is performed on denoised and registered images that are preprocessed in less than 5min (see section 3.1). We compared

the population separation accuracy using manual segmentation of HarP protocol and OPAL segmentation. The robustness and consistency of our automatic segmentation method enable a better group separation between ADNI populations (AD, MCI, NC). Complementary results on the use of automatic segmentations as priors have been also presented. We show that improvements can be obtained without significant increasing of computation time by adding subjects to the training library.

Throughout this paper, we mentioned OPAL high capacities in terms of both segmentation and computation time. With such fast performance, OPAL opens the way for new applications of label fusion segmentation such as integration in visualization software that would highly facilitate the analysis of brain MRI. A web-based tool for on-line remote MRI processing is also a possible application to exploit OPAL capacities. We plan to include OPAL in the next version of volBrain (`http://volbrain.upv.es`).

Finally, in this paper we only applied our method to the hippocampus segmentation, since it is the most studied structure in the Alzheimer's disease context. Nevertheless, the OPAL method can be applied to the segmentation of any anatomical structure. Future research will focus on the extension of the method to the whole brain segmentation as done in [5]. Our preliminary results suggest that this can be done in less than 2 minutes.

6. Conclusion

In this paper, we propose a novel patch-based segmentation method based on an optimized PatchMatch label fusion. Thanks to the low computational burden of our method, we investigated the potential of a new multi-feature and multi-scale framework with late estimator aggregation. The validation of our approach on hippocampus segmentation applied to two different datasets shows that the proposed method produces competitive results compared to the state-of-the-art approaches. Indeed, OPAL obtained the highest median Dice coefficient with a drastically reduced computation time. In addition, OPAL reaches the inter-

expert reliability on both datasets (90% and 89.0% respectively for ICBM and EADC-ADNI datasets). Therefore, OPAL provides automatic segmentations equivalent in terms of Dice coefficient to inter-expert segmentations in less than 2s of processing for the segmentation step. In addition, the volumes segmented by OPAL are highly correlated to the manually segmented volumes. Finally, the accuracy and reproducibility of OPAL enable to better separate ADNI groups (AD, MCI, NC).

Acknowledgments

This study has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the Investments for the future Program IdEx Bordeaux (ANR-10-IDEX-03-02), Cluster of excellence CPU and TRAIL (HR-DTI ANR-10-LABX-57). We also thank Tong Tong and Daniel Rueckert for providing us complete results of the methods proposed in [20], Sonia Tangaro and Marina Boccardi for providing us complete results of the method proposed in [34], and Katherine Gray and Robin Wolz for providing us complete results of the LEAP method proposed in [47]. Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). The ADNI is funded by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics NV, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffmann-La Roche, Schering-Plough, Synarc Inc., as well as nonprofit partners, the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to the ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee

organization is the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by the Spanish grant TIN2013-43457-R from the Ministerio de Economia y competitividad, NIH grants P30AG010129, K01 AG030514 and the Dana Foundation.

References

[1] D. L. Collins, C. J. Holmes, T. M. Peters, and A. C. Evans. Automatic 3-D model-based neuroanatomical segmentation. *Human Brain Mapping*, 3(3):190–208, 1995.

[2] K. Babalola, B. Patenaude, P. Aljabar, J. Schnabel, D. Kennedy, W. Crum, S. Smith, T. Cootes, M. Jenkinson, and D. Rueckert. An evaluation of four automatic methods of segmenting the subcortical structures in the brain. *NeuroImage*, 47(1):1435–1447, 2009.

[3] J. Barnes, J. Foster, R. G. Boyes, T. Pepple, E. K. Moore, J. M. Schott, C. Frost, R. I. Scahill, and N. C. Fox. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. *NeuroImage*, 40(4):1655–1671, 2008.

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

[5] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.

[6] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726–738, 2009.

[7] D. L. Collins and J. C. Pruessner. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *NeuroImage*, 52(4):1355–1366, 2010.

[8] M. J. Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, N. C. Fox, and S. Ourselin. Steps: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. *Medical Image Analysis*, 17(6):671–684, 2013.

[9] R. Wolz, P. Aljabar, D. Rueckert, R. A. Heckemann, and A. Hammers. Segmentation of subcortical structures and the hippocampus in brain MRI using graph-cuts and subject-specific a-priori information. *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI)*, pages 470–473, 2009.

[10] J. Lötjönen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, D. Rueckert, and Alzheimer's Disease Neuroimaging Initiative. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, 49(3):2352–2365, 2010.

[11] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solorzano. Combination strategies in multi-atlas image segmentation: Application to brain MR data. *Medical Imaging*, 28(8):1266–1277, 2009.

[12] A. Khan, N. Cherbuin, W. Wen, K. J. Anstey, P. Sachdev, and M. F. Beg. Optimal weights for local multi-atlas fusion using supervised learning and dynamic information (superdyn): Validation on hippocampus segmentation. *NeuroImage*, 56(1):126–139, 2011.

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

[14] H. Wang, S. R. Das, J. W. Suh, M. Altinay, J. Pluta, C. Craige, B. Avants, P. A. Yushkevich, and Alzheimer's Disease Neuroimaging Initiative. A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation. *NeuroImage*, 55(3):968–985, 2011.

[15] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–954, 2011.

[16] F. Rousseau, P. A Habas, and C. Studholme. A supervised patch-based approach for human brain labeling. *Medical Imaging*, 30(10):1852–1862, 2011.

[17] S. Hu, P. Coupé, J. C. Pruessner, and D. L. Collins. Nonlocal regularization for active appearance model: Application to medial temporal lobe segmentation. *Human brain mapping*, 35(2):377–395, 2014.

[18] L. Wang, F. Shi, G. Li, Y. Gao, W. Lin, J. H. Gilmore, and D. Shen. Segmentation of neonatal brain MR images using patch-driven level sets. *NeuroImage*, 84:141–158, 2014.

[19] S. F. Eskildsen, P. Coupé, V. Fonov, J. V. Manjón, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Østergaard, D. L. Collins, and Alzheimer's Disease Neuroimaging Initiative. BEaST: Brain Extraction based on nonlocal Segmentation Technique. *NeuroImage*, 59(3):2362–2373, 2012.

[20] T. Tong, R. Wolz, P. Coupé, J. V. Hajnal, D. Rueckert, and Alzheimer's Disease Neuroimaging Initiative. Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling. *NeuroImage*, 76:11–23, 2013.

[21] G. Wu, Q. Wang, D. Zhang, F. Nie, H. Huang, and D. Shen. A generative probability model of joint label fusion for multi-atlas based brain segmentation. *Medical Image Analysis*, 18(6):881–890, 2014.

[22] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. PatchMatch: A Randomized Correspondence Algorithm for Structural image Editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), 2009.

[23] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhatia, A. M. Simoes Monteiro de Marvao, T. Dawes, D. O'Regan, and D. Rueckert. Cardiac Image Super-Resolution with Global Correspondence Using Multi-Atlas PatchMatch. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 8151, pages 9–16. 2013.

[24] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein. The Generalized PatchMatch Correspondence Algorithm. In *European Conference on Computer Vision (ECCV)*, volume 6313 of *LNCS*, pages 29–43. 2010.

[25] J. V. Manjón, S. F. Eskildsen, P. Coupé, J. E. Romero, D. L. Collins, and M. Robles. NICE: Non-local Intracranial Cavity Extraction. *International Journal of Biomedical Imaging*, page Article ID 820205, 2014.

[26] G. Wu, M. Kim, G. Sanroma, Q. Wang, B. C. Munsell, D. Shen, and Alzheimer's Disease Neuroimaging Initiative. Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition. *NeuroImage*, 106:34–46, 2015.

[27] C. Wachinger, M. Brennan, G. Sharp, and P. Golland. On the Importance of Location and Features for the Patch-Based Segmentation of Parotid Glands. *Image-Guided Adaptive Radiation Therapy (IGART)*, 2014.

[28] M. Kim, G. Wu, W. Li, L. Wang, Y. D. Son, Z. H. Cho, and D. Shen. Automatic hippocampus segmentation of 7.0 Tesla MR images by combining multiple atlases and auto-context models. *NeuroImage*, 83:335–345, 2013.

[29] W. Bai, W. Shi, C. Ledig, and D. Rueckert. Multi-atlas segmentation with augmented features for cardiac MR images. *Medical Image Analysis*, 19:98–109, 2015.

[30] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *ACM international conference on Multimedia*, pages 399–402, 2005.

[31] C. R. Jack, M. A Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, and C. Ward et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[32] J. C. Mazziotta, A. W. Toga, A. C. Evans, P. Fox, and J. Lancaster. A probabilistic atlas of the human brain: theory and rationale for its development. *NeuroImage*, 2(2):89–101, 1995.

[33] M. Boccardi, M. Bocchetta, L. G. Apostolova, J. Barnes, G. Bartzokis, G. Corbetta, C. DeCarli, M. Firbank, R. Ganzola, and L. Gerritsen et al. Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. *Alzheimer's & Dementia*, 11(2):126–138, 2014.

[34] S. Tangaro, N. Amoroso, M. Boccardi, S. Bruno, A. Chincarini, G. Ferraro, G. B. Frisoni, R. Maglietta, A. Redolfi, L. Rei, A. Tateo, R. Bellotti, and Alzheimers Disease Neuroimaging Initiative. Automated voxel-by-voxel tissue classification for hippocampal segmentation: Methods and validation. *Physica Medica*, 30(8):878–887, 2014.

[35] J. V. Manjón, P. Coupé, L. Martí-Bonmatí, D. L. Collins, and M. Robles. Adaptive non-local means denoising of MR images with spatially varying noise levels. *Journal of Magnetic Resonance Imaging*, 31(1):192–203, 2010.

[36] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushke-

vich, and J. C. Gee. N4ITK: improved N3 bias correction. *Medical Imaging*, 29(6):1310–1320, 2010.

[37] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.

[38] N. Weiskopf, A. Lutti, G. Helms, M. Novak, J. Ashburner, and C. Hutton. Unified segmentation based correction of R1 brain maps for RF transmit field inhomogeneities (UNICORT). *NeuroImage*, 54(3):2116–2124, 2011.

[39] J. V. Manjón, J. Tohka, G. García-Martí, J. Carbonell-Caballero, J. J. Lull, L. Martí-Bonmatí, and M. Robles. Robust MRI brain tissue parameter estimation by multistage outlier rejection. *Magnetic Resonance in Medicine*, 59(4):866–873, 2008.

[40] J. C. Pruessner, L. M. Li, W. Serles, M. Pruessner, D. L. Collins, N. Kabani, S. Lupien, and A. C. Evans. Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. *Cerebral cortex*, 10(4):433–442, 2000.

[41] P. Coupé, J. V. Manjón, E. Gedamu, D. Arnold, M. Robles, and D. L. Collins. Robust Rician noise estimation for MR images. *Medical image analysis*, 14(4):483–493, 2010.

[42] P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. *Medical Imaging*, 27(4):425–441, 2008.

[43] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *Medical Imaging*, 17(1):87–97, 1998.

[44] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of Computer Assisted Tomography*, 18(2):192–205, 1994.

[45] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer. Morphometric analysis of white matter lesions in MR images: method and validation. *Trans Med Imaging*, 13(4):716–724, 1994.

[46] F. Roche, J. Schaerer, S. Gouttard, A. Istace, B. Belaroussi, H. J. Yu, L. Bracoud, C. Pachai, C. DeCarli, and Alzheimer's Disease Neuroimaging Initiative. Accuracy of BMAS Hippocampus Segmentation Using the Harmonized Hippocampal Protocol. *Alzheimer's & Dementia*, 10(4):56, 2014.

[47] K. R. Gray, M. Austin, R. Wolz, K. McLeish, M. Boccardi, G. Frisoni, and D. Hill. Integration of EADC-ADNI Harmonised hippocampus labels into the LEAP automated segmentation technique. *Alzheimer's & Dementia*, 10:555, 2014.

[48] J. Pipitone, M. T. M. Park, J. Winterburn, T. A. Lett, J. P. Lerch, J. C. Pruessner, M. Lepage, A. N. Voineskos, M. Mallar Chakravarty, and Alzheimer's Disease Neuroimaging Initiative. Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *NeuroImage*, 101:494–512, 2014.

[49] B. T. Wyman, D. J. Harvey, K. Crawford, M. A Bernstein, O. Carmichael, P. E Cole, P. K Crane, C. DeCarli, N. C. Fox, and J. L. Gunter et al. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's & Dementia*, 9(3):332–337, 2013.

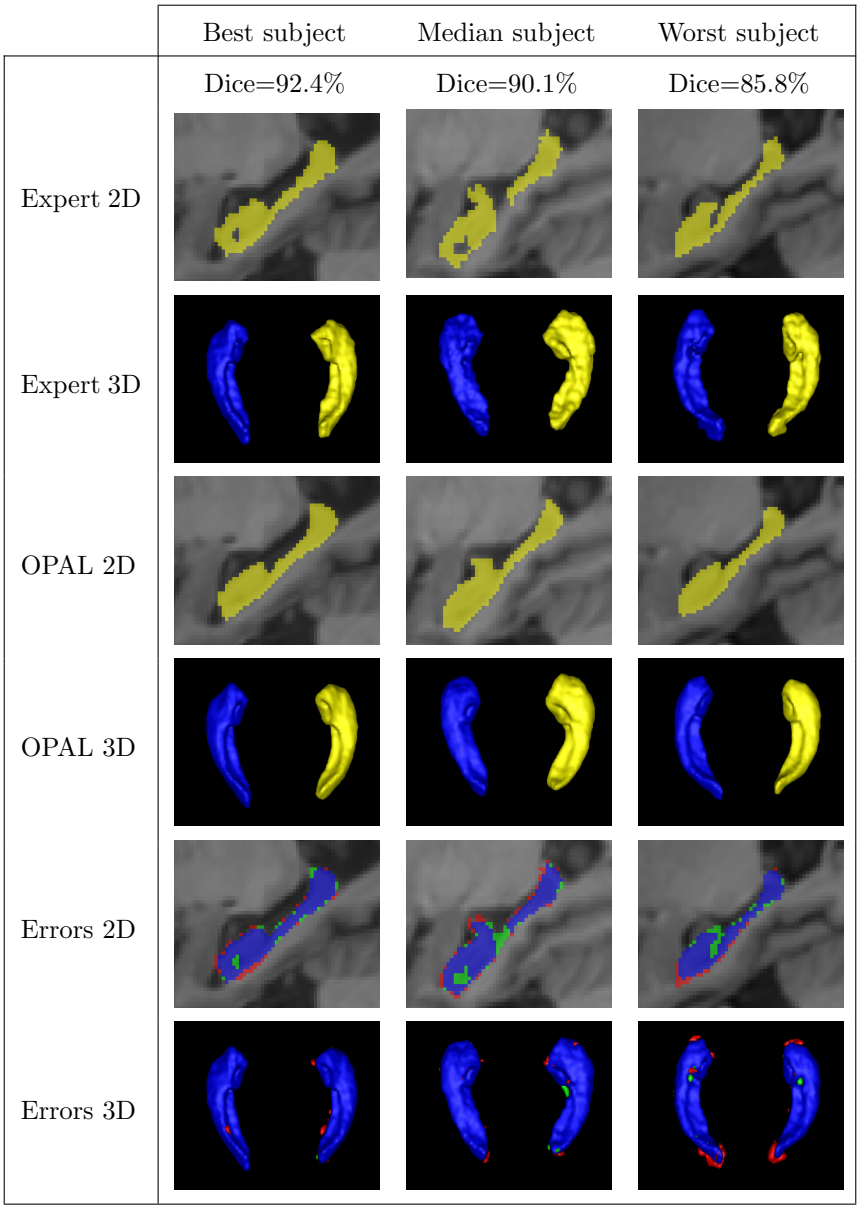


Figure 8: 2D and 3D visualizations of best, median and worst segmented EADC-ADNI subjects computed with default settings. In the fifth and sixth rows, blue voxels are overlapping with the expert segmentation, green voxels are the false positives (segmented by OPAL but not by the expert) and red voxels are the false negatives (segmented by the expert but not by OPAL).