Nonlocal Patch-Based Label Fusion for Hippocampus Segmentation

Pierrick Coupé, Jose Vicente Manjon, Vladimir Fonov, Jens C. Pruessner, Montserrat Robles, Louis D. Collins

Nonlocal Patch-based label fusion for Hippocampus Segmentation

Pierrick Coupé[1], José V. Manjón[2], Vladimir Fonov[1], Jens Pruessner[1], Montserrat Robles[2], D. Louis Collins[1]

[1] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada. University, 3801 University Street, Montreal, Canada H3A 2B4.
[2] Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain

Abstract. Quantitative magnetic resonance analysis often requires accurate, robust and reliable automatic extraction of anatomical structures. Recently, template-warping methods incorporating a label fusion strategy have demonstrated high accuracy in segmenting cerebral structures. In this study, we propose a novel patch-based method using expert segmentation priors to achieve this task. Inspired by recent work in image denoising, the proposed nonlocal patch-based label fusion produces accurate and robust segmentation. During our experiments, the hippocampi of 80 healthy subjects were segmented. The influence on segmentation accuracy of different parameters such as patch size or number of training subjects was also studied. Moreover, a comparison with an appearance-based method and a template-based method was carried out. The highest median kappa value obtained with the proposed method was 0.884, which is competitive compared with recently published methods.

Keywords: hippocampus segmentation, nonlocal means estimator.

1 Introduction

The crucial role of magnetic resonance (MR) imaging has been demonstrated in the detection of pathology, the study of brain organization and in clinical research. However, the vast amount of data produced everyday in clinical settings prevents the use of manual approaches to data analysis. The development of accurate, robust and reliable segmentation techniques for the automatic extraction of anatomical structures is becoming an important challenge in quantitative MR analysis**.** To achieve this task, several automatic methods have been proposed, such as deformable models or region growing [1-2], appearance-based models [3-4] and atlas/template-warping techniques [5-9]. Recently, template-warping techniques that use a library of templates (i.e., MR images with manual expert-based segmentation) have been investigated intensively because of their high accuracy to segment anatomical structures. Barnes et al. [5] proposed to register the most similar template from a library of pre-labeled subjects to segment the hippocampus (HC). However, the use of only one template may result in a biased segmentation. To avoid this problem, it is possible to use several similar

templates [6-7, 9-11]. This requires a label fusion strategy [6, 8-9, 11] to efficiently merge the information derived from the selected templates. In such approaches, two main assumptions are made. First, constraints on structure shape are used implicitly due to the one-to-one correspondence between the voxels of the image to be segmented and those of the warped templates. This presents the advantage of forcing the resulting segmentation to have a similar shape to those of expert-labeled structures in the template library. However, according to the regularization used during registration, some details can be lost and local high variability cannot be captured. Second, label fusion techniques usually assign the same weight to all samples during a vote procedure and consider only the absolute number as a criterion. This approach is sensitive to registration error, since it does not take into account the anatomical relevance of each sample [10]. Therefore, we propose a patch-based scheme with a weighted label fusion, where the weight of each sample is only driven by the similarity of intensity between patches.

In this study, we describe a fully automated patch-based method and the different steps required for its utilization, such as the library construction. Our method is applied to the HC segmentation of healthy subjects. During experiments the influences of different parameters were studied, and a comparison with two other methods was performed. Finally, we discuss further improvements and questions revealed by this new approach.

2 Materials and Methods

Overview

First, the patch library is constructed by removing image acquisition variability and linearly aligning the subjects of the library in a common space. For each voxel of the image to be segmented, the surrounding patch is then compared to the patches contained in the library. Finally, the expert-based information is merged using a nonlocal means approach [12] to produce the final segmentation.

Dataset

The HC dataset consists of 80 subjects selected from 152 T1-weighted (T1w) MR images of young, healthy individuals acquired in the context of the International Consortium for Brain Mapping (ICBM) project. The volunteers were scanned with a 1.5T Philips GyroScan imaging system (1 mm thick slices, TR = 17 ms, TE = 10 ms, flip angle = 30 °, 256 mm field of view). The local ethics committee approved the study and informed consent was obtained from all participants. The 80 selected subjects were manually segmented by an expert. This data subset comprises 39 males and 41 females of similar ages (mean age: 25.09 ± 4.9 years). The resulting segmentations obtained an intra-class reliability coefficient (ICC) of 0.900 for inter- (4 raters) and 0.925 for intra-rater (5 repeats) reliability.

Patch library construction

Denoising: All images in the database were first denoised with the three-dimensional (3D) block-wise nonlocal means filter proposed for MR images in [13].

Inhomogeneity correction: To ensure that each tissue type has the same intensity within a single image, the well-known N3 intensity non-uniformity correction of Sled et al. [14] was used.

Linear registration to stereotaxic space: Each subject was linearly registered to the MNI 152 template into the stereotaxic space using ANIMAL [15].

Intensity normalization: Finally, the intensities of the images were set in [0-100] and were normalized together by using the method proposed in [16]. Since our method involves the matching of a sub-region of anatomical structures based on intensity, the contrast and the luminance information are preserved by performing the global normalization of the entire 3D image.

Search strategy within the library

Initialization mask: Instead of performing the segmentation of the entire image under study, we define an initialization mask around the structure of interest. Many different strategies can be used to propose an accurate initialization, such as the matching of the best subject [5] followed by a morphological dilation of the mask. In this study, we chose a very fast and simple approach that consists in using the union of all the expert segmentations in the training database as the initial mask. In this way, we ensure that the structure is completely included in the mask and demonstrate the robustness of our method to coarse initialization (median Dice kappa of initial mask was around 0.4).

Subject selection: A selection is also performed at the subject level. This strategy is similar to the selection of best subjects in label fusion methods [7]. In our method, we use the sum of the squared difference across the initialization mask instead of using normalized mutual information over the entire image, as suggested in [7]. This strategy was chosen because our patch comparison is based on the L2-norm. Thus, we want to prioritize subjects with similar anatomy as well as similar luminance and contrast. The N closest subjects are finally retained for use during the patch comparison.

Patch pre-selection: As proposed for denoising purposes [13], we pre-select the patches to be compared. In fact, the main part of computational time is dedicated to computing the intensity-based distance between patches. By using simple statistics such as mean or variance, it is possible to discard a priori the most dissimilar patches [13]. In the proposed approach, we use luminance and contrast criteria to achieve this pre-selection. Based on the first and second terms of the well-known structural similarity measure (SSIM) [17], the pre-selection procedure can be written as follows:

$$ ss = \frac{2.\mu_i.\mu_{s,j}}{\mu_i^{\,2} + \mu_{s,j}^{\,2}} \times \frac{2.\sigma_i.\sigma_{s,j}}{\sigma_i^{\,2} + \sigma_{s,j}^{\,2}} , \qquad (1) $$

where μ represents the means, and σ, the standard deviations of the patches centered on voxel x_i (voxel under consideration) and voxel $x_{s,j}$ at location j in the subject s. If the value ss is superior to a given threshold th (0.95 for all experiments), the intensity distance between patches i and j is computed. The patch mean and variance are pre-computed as maps of local means and local variances, thus avoiding multiple computations.

Search area definition: Initially, the nonlocal means denoising filter was proposed as a weighted average of all the pixels in the image, with patch-based similarity used

to assign the weights [12]. For computational reasons, the entire image cannot be used and the number of pixels involved has to be reduced. As for denoising [12-13], we use a limited search area V_i, defined as a cube centered on the voxel x_i under study. Thus, within each subject, we search similar patches in a cubic region around the location under study. This search area can be viewed as the inter-subject variability of the structure of interest in stereotaxic space. This variability can increase for a subject with pathology or according to the structure under consideration.

Nonlocal patch-based label fusion

Nonlocal means estimator: For all voxels x_i of the image to be segmented (included in the initialization mask), the estimation of the final label is based on a weighted label fusion $v(x_i)$ of all labeled samples in the selected library (i.e., inside the search area Vi for the N considered subjects):

$$v(x_i) = \frac{\sum_{s=1}^{N}\sum_{j \in V_i} w(x_i, x_{s,j}) \cdot y_{s,j}}{\sum_{s=1}^{N}\sum_{j \in V_i} w(x_i, x_{s,j})} \, , \qquad (2)$$

where $y_{s,j}$ is the label given by the expert to voxel $x_{s,j}$ at location j in subject s. The weight $w(x_i, x_{s,j})$ is computed as:

$$w(x_i, x_{s,j}) = \begin{cases} \exp^{\frac{-\|P(x_i) - P(x_{s,j})\|_2^2}{h}} & if \quad ss > th \\ 0 & else \end{cases} \, , \qquad (3)$$

where $\|.\|_2$ is the L2-norm, normalized by the number of elements, computed between each intensity of the elements of the patches $P(x_i)$ and $P(x_{s,j})$. If the structure similarity ss between patches is less than th, the weight is not computed and is set directly to zero. Finally, by considering the labels y defined as $\{0,1\}$, the final label $L(x_i)$ is computed as:

$$L(x_i) = \begin{cases} 1 & v(x_i) > 0.5 \\ 0 & v(x_i) < 0.5 \end{cases} \, . \qquad (4)$$

In the event that all patches in the library have ss < th, −1 is returned to indicate that the selected library does not allow a decision to be made.

Local adaptation of h : As usual in estimation problems using a robust function, the tuning of the decay parameter h plays a crucial role. When h is very low, only a few samples are taken into account. When h is very high, all samples tend to have the same weight and the estimation is similar to a classical average. The value of h should depend on the distance between the patch under consideration and the library content. To automatically achieve this local adaptation of h, we propose an estimation of $h(x_i)$, based on the minimal distance between the patch under consideration and the selected subpart of the library:

$$h(x_i) = \arg\min_{x_{s,j}} \|P(x_i) - P(x_{s,j})\|_2 + \varepsilon \, , \qquad (5)$$

where ε is a small constant to ensure numerical stability in case the patch under consideration is contained in the library.

Validation framework

A leave-one-out procedure was performed for the 80 subjects. Dice's kappa was then computed by comparing the expert-based segmentations with the segmentations obtained with our method. The impact of the patch size, search area size and number of training subjects was studied. Moreover, the proposed patch-based method was compared with an appearance-based approach using level-set shape constraints [3], and a template-based technique inspired by the work of Barnes et al. [5] that uses ANIMAL [15] for the nonlinear registration of the best subject. For the appearance-based method, only one modality was used during the process. We used the 79 remaining subjects to construct the training dataset involved in PCA computation. For the template-based method inspired by [5], the best subject was selected using the normalized mutual information, as suggested in [7], and then nonlinearly warped to the subject under study with ANIMAL within a multi-resolution framework until a resolution of 2 mm.

3 Results

The Dice kappa values obtained with the initial standard mask was 0.44, which corresponds to coarse initialization.
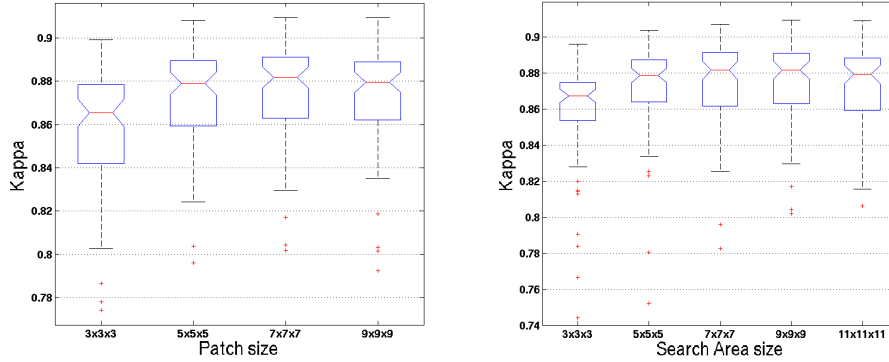

Figure 1: Dice Kappa values according to the patch size (left) and the search area size (right). The results are obtained with 20 training subjects.

Impact of the patch size: First, we studied the impact of patch size on segmentation accuracy. The kappa results are presented at left in Fig. 1. The best median Dice kappa was obtained with a patch size of 7x7x7 voxels for the HC dataset (K = 0.882).

Impact of the search area size: We also studied the impact of the search area size on segmentation accuracy. The kappa results are presented at right in Fig. 1. The best median kappa was obtained with a search area of 9 x 9 x 9 voxels (K = 0.882).

Impact of the number of training subjects: The last important parameter of the proposed method is the number of selected training subjects. During this experiment, segmentation accuracy was studied for N=2 to N=30 selected training subjects. As described previously, the N best training subjects are selected from the 79 remaining ones. The results are presented in Fig. 2. The median kappa value was 0.848 for 2 subjects and 0.884 for 30 subjects. As expected, increasing the number of selected training subjects increased the accuracy of the segmentation.
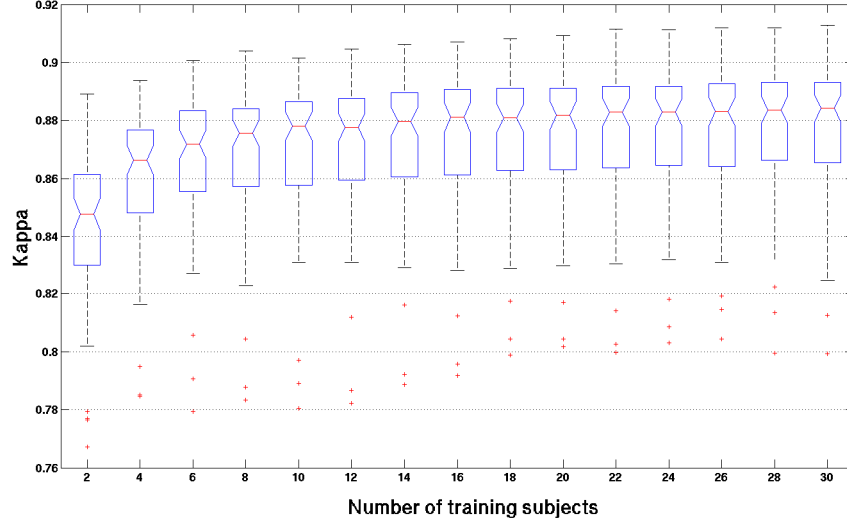
Figure 2: Dice Kappa values according to the number of training subjects obtained with a patch size of 7x7x7 voxels and search area size of 9x9x9 voxels.

Comparison with appearance-based and template-based methods: Finally, the proposed patch-based method was compared with two other methods. Figure 3 presents the kappa values obtained for each method. The results presented for our method were obtained with N = 20. The appearance-based method obtained a median kappa value of 0.800; the best template approach obtained 0.837, whereas the proposed method obtained 0.882. One can note that by using only 2 training subjects (K=0.848) our method already outperforms the two other methods. Figure 4 shows HC segmentations obtained by the methods compared.
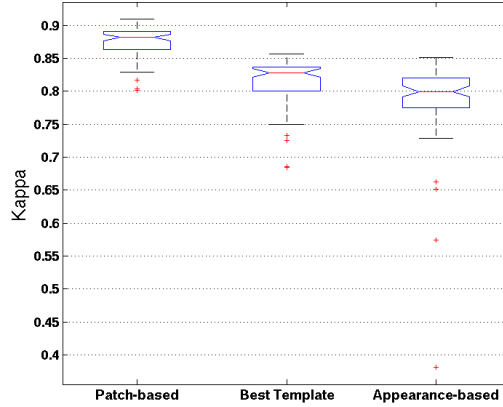

Figure 3: Kappa values obtained by the three methods. The patch-based approach obtained significantly better results compared to the two others methods with a *p-value* << 0.001 in both cases using Kruskal-Wallis tests.

Computational time: The computational time was proportional to the number of subjects; for each subject, around 40 seconds were required. Compared with other

approaches, the appearance-based method [3] took around 1 minute to provide the segmentation of the HC. The best template-based approach inspired by [5] required around 6 minutes to achieve the nonlinear registration of the cropped images already linearly registered into stereotaxic space. However, the comparison of computational time is difficult since our method was coded in MATLAB© and not in C like the other two methods.

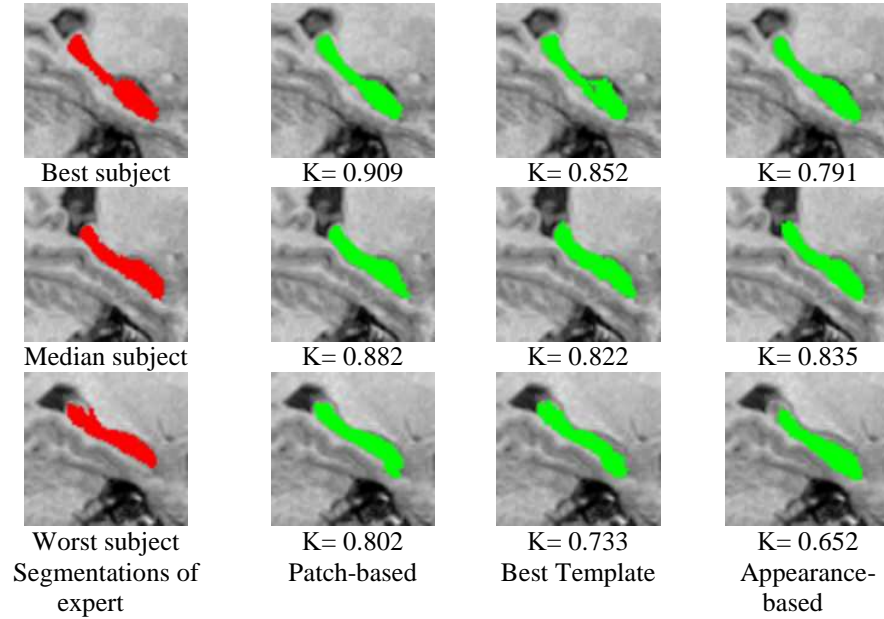|			
Best subject	K= 0.909	K= 0.852	K= 0.791
Median subject	K= 0.882	K= 0.822	K= 0.835
Worst subject	K= 0.802	K= 0.733	K= 0.652
Segmentations of expert	Patch-based	Best Template	Appearance-based

Figure 4: Segmentation results with the three methods compared, for the best (top), a median (middle) and the worst (bottom) subjects obtained with our method.

4   Conclusion

We proposed a novel patch-based approach to automatically segment anatomical structures using expert segmentation priors. Despite its simplicity, the accuracy of the proposed method has been demonstrated within our validation framework. The highest Dice kappa values obtained during experiments were 0.884 for N = 30 training subjects. Moreover, comparison with an appearance-based [3] and a template-based method [5] highlighted the competitive results obtained by the proposed nonlocal patch-based approach. Comparing published methods is always difficult due to differences within the databases used for validation, the studied populations, the quality of expert segmentations and the reported quality metrics. However, tendencies in method evolution and their respective performances can be obtained by studying published results. Recently published results [1, 5, 7] indicated kappa values lower than 0.88. To the best of our knowledge, only the methods based on nonlinear warping of the best templates and involving a label fusion step [9-11] obtained a kappa value equal to or greater than 0.88. Gousias et al. [11] reported a mean kappa of 0.88 with the use of a b-spline-based nonlinear registration on a 2-year-old brain. Lotjonen et al. [10] proposed two intensity-based models to improve

label fusion. With the graph-cut-based method, they obtained a kappa of 0.88, and with the EM-based algorithm, a kappa of 0.885. Collins and Pruessner [9] obtained a median kappa of 0.886 by using nonlinear registration of 11 best templates and a classical voting scheme for label fusion. By comparison, our proposed method offers the main advantages of its simplicity for similar segmentation accuracy (K = 0.884). As a result of the proposed automatic adaptation of the robust function parameter, our approach can be used easily and implemented in a fully automatic manner.

References

1. Chupin, M., et al., Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with Alzheimer's disease. NeuroImage, 2007. 34(3): p. 996-1019.
2. Shen, D., et al., Measuring size and shape of the hippocampus in MR images using a deformable shape model. NeuroImage, 2002. 15(2): p. 422-34.
3. Hu, S. and D.L. Collins, Joint level-set shape modeling and appearance modeling for brain structure segmentation. NeuroImage, 2007. 36(3): p. 672-83.
4. Duchesne, S., J. Pruessner, and D.L. Collins, Appearance-based segmentation of medial temporal lobe structures. NeuroImage, 2002. 17(2): p. 515-31.
5. Barnes, J., et al., A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. NeuroImage, 2008. 40(4): p. 1655-71.
6. Heckemann, R.A., et al., Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage, 2006. 33(1): p. 115-26.
7. Aljabar, P., et al., Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage, 2009. 46(3): p. 726-38.
8. Hammers, A., et al., Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: a proof-of-principle study. NeuroImage, 2007. 36(1): p. 38-47.
9. Collins, D. and J. Pruessner, Towards Accurate, Automatic Segmentation of the Hippocampus and Amygdala from MRI, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009. 2009. p. 592-600.
10. Lotjonen, J.M., et al., Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage, 2010. 49(3): p. 2352-65.
11. Gousias, I.S., et al., Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. NeuroImage, 2008. 40(2): p. 672-84.
12. Buades, A., B. Coll, and J.M. Morel, A non-local algorithm for image denoising. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 2, Proceedings, 2005: p. 60-65.
13. Coupe, P., et al., An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans Med Imaging, 2008. 27(4): p. 425-41.
14. Sled, J.G., A.P. Zijdenbos, and A.C. Evans, A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging, 1998. 17(1): p. 87-97.
15. Collins, D.L., et al., Automatic 3-D model-based neuroanatomical segmentation. Human Brain Mapping, 1995. 3(3): p. 190-208.
16. Nyul, L.G. and J.K. Udupa, Standardizing the MR image intensity scales: making MR intensities have tissue specific meaning. Medical Imaging 2000: Image Display and Visualization, 2000. 1(21): p. 496-504.
17. Wang, Z., et al., Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004. 13(4): p. 600-612.