Simultaneous Segmentation and Grading of Hippocampus for Patient Classification with Alzheimer's Disease

Pierrick Coupé[1], Simon F. Eskildsen[1], José V. Manjón[2], Vladimir Fonov[1], D. Louis Collins[1], and The Alzheimer's Disease Neuroimaging Initiative[*]

[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. Purpose: To propose an innovative approach to better detect Alzheimer's Disease (AD) based on a finer detection of hippocampus (HC) atrophy patterns. Method: In this paper, we propose a new approach to simultaneously perform segmentation and grading of the HC to better capture the patterns of pathology occurring during AD. Based on a patch-based framework, the novel proposed grading measure estimates the similarity of the patch surrounding the voxel under study with all the patches present in different training populations. The training library used during our experiments was composed by 2 populations, 50 Cognitively Normal subjects (CN) and 50 patients with AD. Tests were completed in a leave-one-out framework. Results: First, the evaluation of HC segmentation accuracy yielded a Dice's Kappa of 0.88 for CN and 0.84 for AD. Second, the proposed HC grading enables detection of AD with a success rate of 89%. Finally, a comparison of several biomarkers was investigated using a linear discriminant analysis. Conclusion: Using the volume and the grade of the HC at the same time resulted in an efficient patient classification with a success rate of 90%.

Keywords: hippocampus segmentation, hippocampus grading, patient classification, nonlocal means estimator, Alzheimer's disease.

1 Introduction

The atrophy of medial temporal lobe structures, such as the hippocampus (HC) and entorhinal cortex, is potentially specific and may serve as early biomarkers of

[*] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). 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. ADNI investigators include (complete listing available at www.loni.ucla.edu/ADNI/Collaboration/ADNI Author ship list.pdf).

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 149–157, 2011.

Alzheimer's disease (AD) [1]. In particular, the atrophy of the HC can be used as a marker of AD progression since changes in HC are closely related to changes in cognitive performance of the subject [1]. The evaluation of HC atrophy is usually estimated by volumetric studies on anatomical MRI, requiring a segmentation step that can be very time consuming when done manually. This limitation can be overcome by using automatic segmentation methods [2-5]. However, despite the high segmentation accuracy of these HC segmentation approaches, using only the HC volume enables a separation between AD and cognitively normal (CN) subjects with a success rate around 72-74% [6]. This limited capability to classify AD patients by using the HC volume only may be due to a simplification of the complex hippocampal atrophy patterns to a volume changing measurement. Recently, several shape analysis methods have been proposed [7-8] to capture detailed patterns of change in order to obtain a more accurate classification. These approaches provide a slightly better classification rate of around 77% [6].

Inspired by work in image denoising [9], a new nonlocal patch-based label fusion method has recently been proposed to segment anatomical structures [5]. By taking advantage of the redundancy of information present within the subject's image, as well as the redundancy across the training subjects, the patch-based nonlocal means scheme enables robust use of a large number of samples during estimation. In [5], this approach has been applied to label fusion for the segmentation of anatomical structures. We propose an extension of this patch-based segmentation method in order to evaluate the similarity (in the nonlocal means sense) of the intensity content of one MRI compared to several training populations. By using training populations with different pathological status (e.g., CN subjects and patients with AD), a nonlocal means estimator is used to evaluate the proximity (i.e., the grade or the degree of atrophy in case of AD) of each voxel of the MRI under study compared to the training populations. Since the grade estimation and the label fusion steps require the same patch comparison, simultaneous segmentation and grading of HC can be achieved in one pass. In the proposed approach, the nonlocal patch-based comparison is used to efficiently fuse the HC segmentations of MRI in a training database and at the same to aggregate the pathological status of the populations constituting the training database. Finally, the average grading value obtained over the segmented HC is proposed as a new biomarker to estimate the pathological status of the subject under study. The contributions of the paper are: *i)* the introduction of an innovative approach to better characterize the patterns of pathology (e.g., atrophy) in AD through the new concept of HC grading, *ii)* the presentation of a method to automatically and simultaneously perform the segmentation and the grading of HC, and *iii)* the demonstration that the proposed approach can be used as a novel biomarker to efficiently achieve patient classification in the context of AD.

2 Materials and Methods

2.1 Dataset and Preprocessing

In this study, the ADNI database (www.loni.ucla.edu/ADNI) was used to validate the proposed approach. This database contains both 1.5T and 3.0T T1-w MRI scans. For

our experiments, we randomly selected 120 MRI scans, 60 1.5T MRI baseline scans of CN subjects and 60 1.5T MRI baseline scans of patients with AD. All the selected images were preprocessed as follows: 1) correction of inhomogeneities using N3 [10], 2) registration to the stereotaxic space using a linear transform to the ICBM152 template (1x1x1 mm³ voxel size) [11] and 3) cross-normalization of the MRI intensity using the method proposed in [12]. After preprocessing, all the MRIs are coarsely aligned (linear registration), tissue intensities are homogeneous within each MRI volume (inhomogeneity correction) and across the training database (intensity normalization). From the 120 processed MRI scans, 20 scans (10 CN and 10 AD) were randomly selected to be used as **seed dataset**. The left and right hippocampi of this **seed dataset** were then manually segmented by an expert at our centre. The manual segmentations of the **seed dataset** were propagated to the 100 remaining scans constituting our **test dataset**. After segmentation propagation using [5], the **test dataset** was composed of 100 MRI (50 CN subjects and 50 patients with AD) with their corresponding automatic segmentations.

2.2 Method Overview

In nonlocal means-based approaches [9], the patch $P(x_i)$ surrounding the voxel x_i under study is compared with all the patches $P(x_j)$ of the image Ω whatever their spatial distance to $P(x_i)$ (it is the meaning of the term "nonlocal"). According to the patch similarity between $P(x_i)$ and $P(x_j)$, estimating by the Sum Squared Difference (SSD) measure, each patch receives a weight $w(x_i, x_j)$:

$$w(x_i, x_j) = e^{\frac{-\|P(x_i)-P(x_j)\|_2^2}{h^2}}$$

where $\|.\|_2$ is the L2-norm computed between each intensity of the elements of the patches $P(x_i)$ and $P(x_{s,j})$, and h is the smoothing parameter of the weighting function. This weighting function is designed to give a weight close to 1 when the SSD is close to zero and a weight close to zero with the SSD is high. Finally, all the intensities $u(x_j)$ of the central voxels of the patches $P(x_j)$ are aggregated through a weighted average using the weights $w(x_i, x_j)$. In this way, the denoised intensity $\hat{u}(x_i)$ of the voxel x_i can be efficiently estimated:

$$\hat{u}(x_i) = \frac{\sum_{j\in\Omega} w(x_i, x_j) u(x_j)}{\sum_{j\in\Omega} w(x_i, x_j)}$$

In [5], we introduced this estimator in the context of segmentation by averaging labels instead of intensities. By using a training library of N subjects, whose segmentations of structures are known, the weighted label fusion is estimated as follows:

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

where $l(x_{s,j})$ is the label (i.e., 0 for background and 1 for structure) given by the expert to the voxel $x_{s,j}$ at location j in training subject s. It has been shown that the nonlocal means estimator $v(x_i)$ provides a robust estimation of the expected label at x_i [5]. With

a label set of {0,1}, voxels with value $v(x_i) \geq 0.5$ are considered as belonging to HC and the remaining voxels as background.

In this paper, we propose to extend it to efficiently aggregate pathological status in order to estimate the proximity (in the nonlocal means sense) of each voxel compared to both populations constituting the training library. To do that, we introduce the new concept of patch-based grading that reflects the similarity of the patch surrounding the voxel under study with all the patches present in the different training populations. In this way, the neighborhood information is used to robustly drive the search of anatomical patterns that are specific to a given subset of the training library. When the training populations include data from subsets of patients with different stages of the pathology progression, this approach provides an estimation of the grade (i.e., degree of atrophy in case of AD) for each voxel:

$$g(x_i) = \frac{\sum_{s=1}^{N} \sum_{j \in \Omega} w(x_i, x_{s,j}).p_s}{\sum_{s=1}^{N} \sum_{j \in \Omega} w(x_i, x_{s,j})}$$

where p_s is the pathological status of the training subject s. In our case, $p_s=-1$ was used for AD status and $p_s=1$ for CN status. A negative grading value (respectively, a positive grading value) $g(x_i)$ indicates that the neighborhood surrounding x_i is more characteristic of AD than CN (respectively, of CN than AD). The absolute value $|g(x_i)|$ provides the confidence given to the grade estimation. When $|g(x_i)|$ is close to zero, the method indicates that the patch under study is similarly present in both populations and thus is not specific to one of the compared populations and provides little discriminatory information. When $|g(x_i)|$ is close to 1, the method detects a high proximity of the patch under study with the patches present in one of the training population and not in the other. Finally, for each subject, an average grading value is computed over all voxels in the estimated segmentation (i.e., for all x_i with $v(x_i) \geq 0.5$) by assigning the same weight to the left and right HC (i.e., $\overline{g} = (\overline{g}_{left} + \overline{g}_{right})/2$). During all our experiments, the default parameters proposed in [5] have been used. The patch size was fixed to 7x7x7 voxels and the search window of similar patches has been limited within a restricted volume of 9x9x9 voxels for computational reasons (i.e., Ω is replaced by a cubic volume V_i centered on x_i). Finally, the smoothing parameter h^2 was locally set as the minimal SSD found between the patch under study and all the patches in the training library as proposed in [5].

2.3 Validation Framework

<u>Segmentation accuracy validation:</u> In order to evaluate the segmentation accuracy of the method proposed in [5] on patients with AD, we first perform a leave-one-out procedure on the 20 subjects with manual segmentation composing the **seed dataset**. The $N=16$ closest training subjects (in the SSD sense, see [5] for details) were equally selected within both populations (i.e., 8 AD and 8 CN). The Dice's kappa was then computed by comparing the expert-based segmentation, used as gold standard, and the segmentation obtained automatically. This first validation is used to support the fact that the segmentation propagation over the 100 subjects in our **test dataset** from the 20 subjects in our **seed dataset** is done in an accurate manner.

Grading validation: After the segmentation propagation step, a leave-one-out procedure is performed over the 100 subjects of the **test dataset**. For each subject, the *N* closest training subjects are selected equally in both populations. This is done to ensure that the size of the *"patch pool"* from AD population is coarsely similar to the size of the *"patch pool"* from CN population. To save computational time, *N* is automatically adjusted according to the obtained $\overline{g}$. In the first iteration, *N=20* (10 CN and 10 AD). If the resulting $|\overline{g}| < 0.1$ (i.e., the confidence in the obtained grade is low), the size of the used training library is increased by 20 to *N=40* (20 CN and 20 AD). This process is repeated until $|\overline{g}| > 0.1$ or *N>80.* The sign of the final grading value is used to estimate the pathological status of the testing subjects. Finally, the success rate of the patient classification is provided to demonstrate the robustness of the proposed new biomarker.

Comparison of biomarkers for patient classification with AD: The last part of our validation framework is the comparison of two biomarkers (HC volume and HC grade) and the investigation of their combination. The segmentations obtained at the same time as the grading were used to obtain the HC volume for each of the subjects in the **test dataset**. Through a leave-one-out procedure, each subject was classified by using optimal boundary separating both populations. This optimal boundary was obtained by performing a linear discriminant analysis over the 99 remaining subjects. This approach was applied to volume-based classification, grade-based classification and the combination of both volume and grade. The success rate (SR), the specificity (SPE), the sensitivity (SEN), the positive predictive value (PPV) and negative predictive value (NPV) are presented for each of the tested biomarkers (see [6] for details on these quality metrics).

3 Results

Table 1 shows the segmentation accuracy obtained on the **seed dataset** by using *N=16* training subjects (8 CN and 8 AD). For the CN population, the median Dice's Kappa was similar to the Dice's Kappa presented in [5] on healthy young subjects from the ICBM database, which demonstrates the robustness of the segmentation method. A lower median Dice's Kappa value was obtained for the AD population. A median Dice's Kappa value superior to 0.8 indicates a high correlation between manual and automatic segmentations, and a median Dice's Kappa value superior 0.88 is similar to the highest published values in literature [3-4]. The difference between populations might come from two sources. First, the higher anatomy variability of patients with AD makes the segmentation more difficult and may require a larger training library. Second, the smaller HC volumes of patients with AD, due to the HC atrophy, can negatively bias the Dice's Kappa index measure. Finally, these results indicate that a similar accuracy can be expected during the segmentation propagation step to the 100 subjects of the **test dataset**.

Table 1. Median Dice's Kappa values (with the standard deviation) obtained on the **seed dataset** composed of 20 MRI (10 CN and 10 AD) with manual segmentations.

Median Dice's Kappa (standard deviation)	*Left HC*	*Right HC*	*Both HC*
CN population	**0.891** (0.035)	**0.866** (0.038)	**0.883** (0.037)
AD population	**0.830** (0.042)	**0.858** (0.035)	**0.838** (0.038)

Figure 1 shows the final grading values for the 100 subjects of the **test dataset**. In the perfect case, the 50 first subjects (CN) should have positive average grading values and the 50 last (AD) should have negative average grading values. As shown in the graph, the success rate of the classification was 89% (5 false positive CN and 6 false negative AD). Figure 1 also presents the size of the used training library for each of the testing subjects. Most of the test subjects were classified by using only *N=20* training subjects. Around 5% of test subjects seem to require larger training library (i.e., *N>80*) since at the end of the procedure $|\overline{g}|$ is still inferior to 0.1.

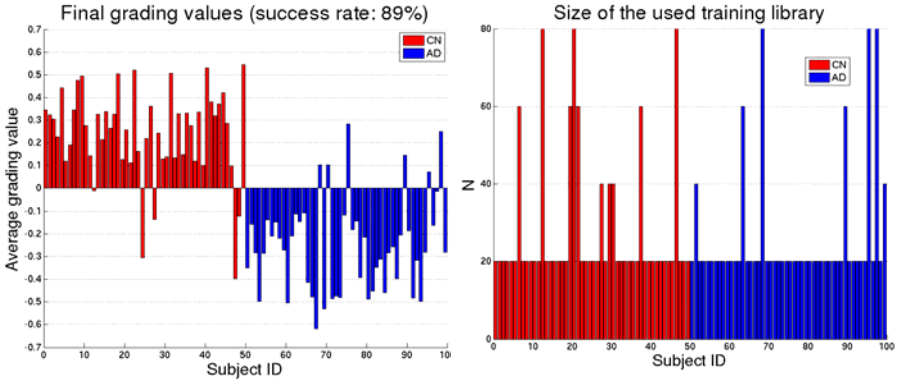

Fig. 1. Left: the final average grading values obtained for the **test dataset**. Right: the used size of training library (i.e., *N*) for all the testing subjects.

Figure 2 shows the grading maps obtained for 2 test subjects (1 CN and 1 AD). The corresponding average grading values and the estimated volumes are also provided for left and right HC. While the volume of HC is similar for these 2 subjects, and thus does not allow an efficient patient classification, their grading values provide a useful indication on their pathological status. Visually, the CN subject clearly appears closer to the CN population (mainly red color related to values close to 1) while the AD patient is visually closer to the AD population (mainly purple and black colors related to values close to -1). Finally, Fig. 2 also provides a visual assessment of the quality of the segmentation propagation on the **test dataset**. For a given subject, the segmentation and the grading maps were obtained in less than 5 minutes using a single core of an Intel Core 2 Quad Q6600 processor at 2.4 GHz with *N=20*.

Table 2 presents the results of the patient classification for the different biomarkers under consideration. These results clearly demonstrate the advantage of using the grading approach (89% of success rate) compared to the classical volumetric approach (78% of success rate). The SEN, SPE, PPV and NPV obtained by our grading approach were higher than the ten methods compared in [6] involving Voxel-Based Morphometry (VBM), cortical thickness, HC volume and HC shape. The higher SR of our volumetric approach compared to the results presented in [6] might come from differences in the test dataset used here or due to a higher accuracy and consistency of the segmentation method used compared to [2]. It is also interesting to note that the optimal boundaries found by linear discriminant analysis provided similar results as using 0 as threshold value as in the previous experiment (see Fig 1.). Finally, using the volume and the average grade of the HC simultaneously provides a very high success rate of 90%.

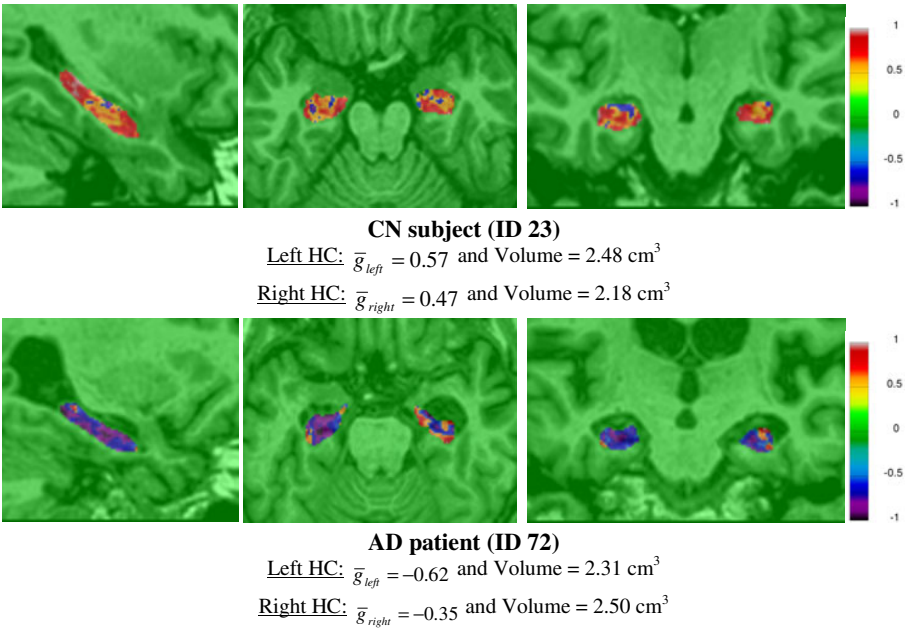

Fig. 2. Top: the obtained grading map for one CN subject (ID 23). Bottom: the obtained grading map for one AD patient (ID 72). The slices of both subjects have the same position in the stereotaxic space. Red color indicates a grading close to 1 (i.e., CN) and black color indicates a grading close to -1 (i.e., AD).

Table 2. Results of the patient classification (AD vs CN) for the different biomarkers under investigation. These results were obtained by using linear discriminant analysis through a leave-one-out procedure on the test dataset.

AD vs. CN	*SR*	*SEN*	*SPE*	*PPV*	*NPV*
HC volume	78%	72%	84%	82%	75%
HC grading	89%	86%	92%	91%	87%
HC volume and grading	90%	88%	92%	92%	88%

4 Conclusion

In this paper, a new method is proposed to robustly detect the hippocampal atrophy patterns accruing during AD. Based on a nonlocal means estimation framework, the proposed novel grading measure (i.e., the atrophy degree in AD context) enables an accurate distinction between CN subjects and patients with AD leading to a success rate of 89% when used alone, and 90% when combined with HC volume. These results are competitive compared to the AD detection performance of VBM, cortical thickness, HC volume and HC shape methods extensively compared in [6]. In contrast to these approaches, our method has the advantage of simplicity (it can be coded in few hundred lines of code), low computational cost (does not required non-rigid registration), robustness of the process (all the subjects get final grading maps) and the possibility to achieve *individual* classifications based on a single time point contrary to *group* classification or longitudinal studies. These first results are promising and indicate that the new HC grading approach could be a useful biomarker to efficiently detect AD. Further work will investigate the possibility to discriminate population of patients with Mild Cognitive Impairment (MCI) compared to AD or CN.

References

1. Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M.: The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology 6(2), 67–77 (2010)
2. Chupin, M., Hammers, A., Liu, R.S., Colliot, O., Burdett, J., Bardinet, E., Duncan, J.S., Garnero, L., Lemieux, L.: Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. Neuroimage 46(3), 749–761 (2009)
3. Collins, D.L., Pruessner, J.C.: 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)
4. Lotjonen, J.M., Wolz, R., Koikkalainen, J.R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D.: Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage 49(3), 2352–2365 (2010)
5. Coupe, P., Manjon, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54(2), 940–954 (2011)
6. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehericy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O.: Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. Neuroimage (2010)
7. Gerardin, E., Chetelat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehericy, S., Garnero, L., Eustache, F., Colliot, O.: Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47(4), 1476–1486 (2009)
8. Csernansky, J.G., Wang, L., Swank, J., Miller, J.P., Gado, M., McKeel, D., Miller, M.I., Morris, J.C.: Preclinical detection of Alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. Neuroimage 25(3), 783–792 (2005)

9. Coupe, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., Barillot, C.: An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans. Med. Imaging 27(4), 425–441 (2008)
10. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17(1), 87–97 (1998)
11. Collins, D.L., Holmes, C.J., Peters, T.M., Evans, A.C.: Automatic 3-D model-based neuroanatomical segmentation. Human Brain Mapping 3(3), 190–208 (1995)
12. Nyul, L.G., Udupa, J.K.: Standardizing the MR image intensity scales: making MR intensities have tissue specific meaning. Medical Imaging 2000: Image Display and Visualization 1(21), 496–504 (2000)