Gradient based Reliability Maps for ACM based Segmentation of Hippocampus

Dimitrios Zarpalas, Polyxeni Gkontra, Petros Daras, *Senior Member, IEEE,* and Nicos Maglaveras, *Senior Member, IEEE*

***Abstract*—Automatic segmentation of deep brain structures, such as the hippocampus (HC), in MR images has attracted considerable scientific attention due to the widespread use of MRI and to the principal role of some structures in various mental disorders. In the literature, there exists a substantial amount of work relying on deformable models incorporating prior knowledge about structures' anatomy and shape information. However, shape priors capture global shape characteristics and thus fail to model boundaries of varying properties; hippocampus' boundaries present rich, poor and missing gradient regions. On top of that, shape prior knowledge is blended with image information in the evolution process, through global weighting of the two terms, again neglecting the spatially varying boundary properties, causing segmentation faults. An innovative method is hereby presented that aims to achieve highly accurate HC segmentation in MR images, based on the modeling of boundary properties at each anatomical location and the inclusion of appropriate image information for each of those, within an ACM framework. Hence, blending of image information and prior knowledge is based on a local weighting map, which mixes gradient information, regional and whole brain statistical information with a multi-atlas based spatial distribution map of the structure's labels. Experimental results on three different datasets demonstrate the efficacy and accuracy of the proposed method.**

***Index Terms*—Hippocampus segmentation, brain MRI, ACM, prior knowledge, local weighting scheme, multi-atlas.**

I. Introduction

Accurate and reliable segmentation of medial temporal lobe structures, such as the hippocampus (HC), from MR images is considered a key requirement for the assessment, treatment and follow-up of various mental disorders [24], including Major Depressive Disorder (MDD), Post-Traumatic Stress Disorder (PTSD), schizophrenia (SD), Alzheimer's Disease, Bipolar disorder (BD), etc. HC is heavily investigated, primarily due its role in memory and contextualization. It has been found to have a key role in the neural mechanisms of major psychiatric diseases [12], [36]. Scientific literature on this topic underlines the involvement of HC in the pathogenesis of SD and BD [49], supporting that altered volume and connectivity of these specific areas may represent a specific endophenotype [10]. Volume deficits of HC have been reported also in first-episode SD patients and in some non-psychotic relatives of SD probands [47].

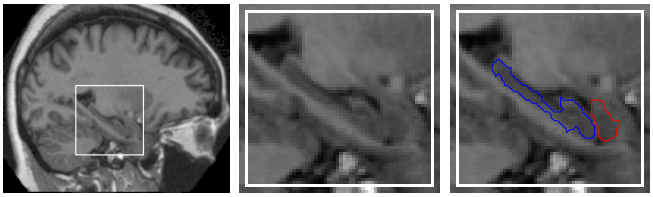


Fig. 1. Sagittal slice of a brain MRI. The white box encloses the HC-amygdala (AG) complex, while the blue and red contours depict the HC and AG boundaries respectively.

Morphological analysis and shape comparisons of HC from healthy and diseased subjects would indicate abnormal deformations, thus leading to possible biomarker identification, disease prognosis and diagnosis, and optimum treatment identification. Automatic segmentation offers reasonable promises, but requires overcoming the inherent difficulties of medical imaging: noise, limited resolution and partial volume effect, resulting in weak boundaries between neighboring structures, especially when they are of the same tissue type (i.e. gray or white matter). Such an example is the case of the hippocampus-amygdala complex, where the imaging resolution is not sufficient to depict the border between them, as Fig. 1 shows. Because of the above and its morphology, HC is considered as a very challenging structure to be segmented, given also that it is among the structures with the lower segmentation accuracy reported in various works (e.g. [4], [9], [26], [35], [43]). Based on the aforementioned, the medical importance of HC in neurodegeneration and the challenge in segmenting it, a lot of works ([17], [18], [19], [23], [31], [34], [38]), and workshops [1], focus only on the accurate and automatic HC segmentation.

Previous work

Various methods for automatic segmentation of challenging brain structures have been proposed so far. These methods are broadly divided into three major categories: i) the atlas or multi-atlas techniques, ii) methods based on deformable models, known as Active Shape Models (ASMs) and Active Contour Models (ACMs) depending on whether they incorporate shape-prior information or not, respectively, and iii) the Active Appearance Models (AAMs), an extension of ASMs. There are methods, not directly classified on these categories, such as [48] that uses a tree of local feature based classifiers to assign a label to each voxel and then apply a shape prior on the results, and [38] where graph cuts are utilized to minimize

Manuscript submitted in July, revised in September and October, and accepted in Novemeber 2013.

D. Zarpalas, P. Gkontra, and P. Daras are with the Information Technologies Institute, Centre for Research and Technology Hellas, Greece. e-mail: {zarpalas, gkontra, daras}@iti.gr

D. Zarpalas and N. Maglaveras are with the Laboratory of Medical Informatics, the Medical School, Aristotle University of Thessaloniki, Greece. email: nicmag@med.auth.gr

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

an energy term that includes both intensity and prior terms, which are based on voxel classification and atlas registration.

Atlas-based methods are well-established in medical image segmentation. Given one or multiple atlases, a segmentation of a new, test image can be produced, through label propagation. Non-rigid registration of the atlas(es) at the target image is performed according to some similarity measure. Segmentation is subsequently performed by applying this transformation on the labeled image(s) with the use of the calculated warp field. In the multi-atlas case, there is the extra step of combining the transformed labeled images (usually through a similarity score), to offer the test image's labels. There are various atlas-based approaches available, [28], [4], [34], [19], [35], [5] whose differences are basically on how to evaluate the accurate registration and on how to perform the fusion of the multiple atlases. A recent enhancement of the label propagation concept, in an effort to avoid the inherent computational cost of the non-rigid registrations, is offered by [23], [43] through patch-based approaches for fusing the label images. Other attempts include fusing the multi-atlas concept with intensity classification and nearest neighbor connectivity [44], with intensity modeling [39], or with multi-scale algorithms that use graph representation [3]. The recent MICCAI 2012 "Workshop on Multi-Atlas Labeling" [2], offered an updated evaluation of multi-atlas techniques for the task of brain segmentation, primarily focusing on variations of the label fusion task. As it appears, for the task of HC segmentation the winning concepts are the joint label fusion, especially combined with corrective learning, as proposed by Wang et al. [51], and the non-local STAPLE proposed by Asman et al. [6]. Detailed evaluation results on this challenge and the outcomes of the proposed method on it, are given in the experimental section of this paper.

ACMs are based on the evolution of a curve or surface according to the intensities' statistical information (region-based ACM) [16] or the image gradients (edge-based ACM) [15] in order to divide the image into meaningful parts. Level-set based ACMs have become very popular, as they can overcome numerical instabilities, while are capable of handling topological changes during evolution. Despite their popularity, by being solely dependent on information extracted from the image, ACMs have been proven insufficient for challenging applications [13]. To overcome this, prior knowledge restrictions can be incorporated. In [17], [18] prior knowledge is structure specific (based on anatomical descriptions of topology, position, distances and their relationships), defined by neuroanatomical landmarks in a training set. This anatomical knowledge is then modeled in the energy guiding the deformation process. In other ASM based approaches, a statistical prior on shape variations from a training set is built by means of Principal Components Analysis (PCA) and incorporated into the segmentation framework to restrain the evolving contour. The first attempt was reported in [20], where shapes were represented by point distribution models, while in [37] signed distance functions were utilized in order to be more robust to misalignments. Later this concept was also extended in [13], where the shape prior was integrated with the Mumford-Shah functional. However, the PCA based statistical prior used in these works imposes global shape constraints. In [52], apart from the shape prior, a neighborhood prior is also modeled, to accommodate the influence between neighboring structures. Readers are referred to [29] for a more thorough analysis on ways for statistical shape modeling of prior knowledge.

ASMs have further been extended to AAMs, which were firstly introduced by Edwards et al. [25] and later extended by Cootes et al. [21], [22]. The concept of AAMs is to model not only the shape prior, but also a texture prior. PCA is used to construct the linear subspaces that model the variation of both shape and texture information in a given population. Segmentation is performed by finding the optimal projection parameters, through matching the synthesized image produced by those parameters, with the test image. Recent advances in AAM-based techniques include multi-band AAMs, where the appearance of derived features is used apart from intensity [46], AAM modeling in combination with patch-based label fusion [30], and the use of level-sets [31], [32], which help to overcome the shortcomings of landmark-based evolution. The major advantage of AAMs is the very light computational cost, however it has been argued in [8] that AAM by being a local search technique requires good initialization, and thus the authors propose the inclusion of a graph-based matching to improve the initialization stage of the algorithm. A more detailed discussion on AAMs can be found in [27].

The aforementioned techniques, and their descendants, primarily focus on different ways of capturing prior knowledge and incorporating it into the segmentation framework. Furthermore, modeling of the varying properties of a structure's boundary is of significant importance. Hippocampus' boundaries present rich, poor and missing gradient regions. Appropriate modeling of such knowledge can be beneficial for its segmentation. Moreover, in most approaches prior knowledge is blended with image information through global weighting of the two terms, causing segmentation faults. In an effort to model the boundary's varying properties, we have proposed in [53] a learning technique to recognize which anatomical locations of the boundaries offer sufficient image information, and to construct a local weighting map, called Gradient Distribution on Boundary (GDB). This map serves for appropriate blending of image and prior information (prior information is modeled through a spatial distribution map of the labels given a training set) during an ACM evolution, i.e. GDB defines at a voxel level, where and at which extent image and/or prior information should be trusted. This concept was recently extended in [54], where GDB became adaptive (AGDB), based on the ACM evolving contour.

In this paper, we tried to make GDB more generic, in order to avoid continuously registering it to the evolving contour, as AGDB does, yet without sacrificing its contribution to the segmentation task. Hence, we extended the nature of GDB; initially, GDB was the local balancer between two energy terms, but the inclusion of more than two types of energy terms led us to construct a three-phase GDB (3GDB), having one phase for the strong edge boundary parts, a second one for the blurred/noisy boundary parts, and a third one for the missing edge boundary parts. In short, the proposed method is an ACM method on top of the multi-atlas concept. It

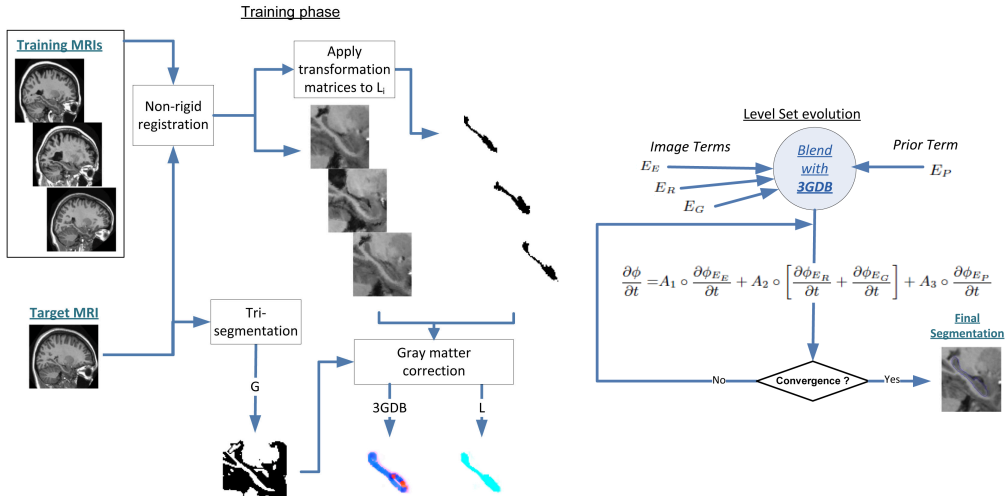


Fig. 2. Overall diagram of the proposed method; the training phase presented on the left can be roughly divided into three steps (i) non-rigid registration of the training MRIs to the target MRI and application of the resulting transformation to the corresponding labeled training images (L_i), (ii) correction of the registered atlases by the use of the target's estimated gray matter produced by a tri-segmentation algorithm (G), and (iii) construction of the 3GDB and the Spatial Distribution Map (L), which incorporate prior knowledge about the varying HC boundary properties and the spatial distribution of hippocampal labels, respectively, using the corrected atlases. The segmentation procedure presented on the right refers to the level-set evolution step. Region information, gradient information, whole brain statistics and prior knowledge about the distribution of labels are blended into a common ACM scheme by the use of the 3GDB in order to drive the level-set evolution.

extracts and models prior information in a twofold way, i.e. a spatial distribution map and a boundary properties map that can both be straightforwardly incorporated into a well defined ACM framework. 3GDB tries to optimally blend three different types of image information with the prior information. More precisely, the main contributions offered are: i) the incorporation of three different types of image information in an effort to better describe the various properties of the desired structures, i.e. whole brain and regional statistical information, and gradient information, ii) a multi-atlas based prior knowledge corrected by the test image, to both initiate the active contour and to incorporate anatomical information into the segmentation procedure, iii) a novel way of capturing prior knowledge about the varying boundary properties of the structures, i.e. the development of 3GDB, and iv) a new dataset of manual segmentations is offered to the rest of the community, in an effort to promote common datasets for experimentation and fair comparisons among the existing methods.

The rest of the paper is organized as follows: section II describes the proposed framework, and more specifically subsection II-C describes how 3GDB is constructed, subsection II-D how it is used to drive the contour evolution, while subsection II-E describes the four energy terms used. Experimental results and comparisons are presented in section III, and finally conclusions are drawn in section IV.

II. Proposed Method

Fig. 2 depicts the overall diagram and the conceptual flow of the proposed methodology. In a nutshell, the proposed method is an ACM method, whose energy to be minimized consists of four energy terms blended by means of 3GDB, on top of the multi-atlas concept. Thus, it starts with the multi-atlas methodology to build prior knowledge, construct 3DGB and then an ACM stage follows, that takes into account the varying boundary properties and both local and global image information, in order to refine the multi-atlas output.

A. Multi-atlas

Consider a set of $i = 1, ..., N$ training images and their corresponding labeled images $L_i, i = 1, ..., N$, where $L_i(\mathbf{v}) = 1$ with $\mathbf{v} = (x, y, z)$ denoting the coordinates of a voxel, for voxels that belong to HC and $L_i(\mathbf{v}) = 0$, otherwise. Each image in the training set is considered to be an atlas and is non-rigidly registered to the test image I. Registration is performed using the symmetric normalization methodology (SyN) [7], from the ANTs toolkit, which is also used to calculate the similarity metric s_i, which is the cross correlation value between the registered anatomical image and the target image. For the calculation of the similarity metric the whole MR images are used rather than a region of interest.

In order to accommodate for registration errors and improve the reliability of the registered atlases, statistics of the whole brain, regarding the tissue distribution of the target image are used. Towards this, a binary map (G) of the gray matter, with 1 indicating a gray matter voxel, is produced by the use of a tissue segmentation algorithm (tri-segmentation). G is referred to as target's estimated gray matter in the rest of the document. The tri-segmentation, classifies each voxel into the three tissue types: cerebrospinal fluid, gray matter and white matter. For this purpose, skull-stripped images, produced by means of BET [45], were used as input to the FAST software tool. Both tools are part of the FSL software suite[1]. The segmentation algorithm used by FAST is based on a hidden Markov random field (HMRF) model, which is combined with an expectation-maximization (EM) algorithm in order to solve the maximum likelihood estimation of the model parameters [55]. Considering HC as a gray matter structure, G is used to exclude non-gray matter parts from the registered atlases. This way, the atlases $AtlasL_i$ are produced, which are warped to the space of the target image and corrected by means of G.

Following registration and correction of all label images of the training set, the multi-atlas based Signed Distribution Map

[1] http://www.fmrib.ox.ac.uk/fsl/

(SDM) is calculated through a weighted averaging and stored in L as:

$$L = \sum_{i=1,..,N} s_i \cdot AtlasL_i \quad (1)$$

where all s_i have been normalized so that $\sum_{i=1,..,N} s_i = 1$. Thus, L offers the labels' distribution, i.e. how likely a voxel $\mathbf{v}$ belongs to the desired structure.

B. 3GDB based ACM on top of multi-atlas

The ACM energy to be minimized contains four energy terms in total. The first three are image terms, which contain information extracted from the test image itself, while the fourth term is the prior term, containing information extracted from a training set.

The first image term is the edge term (E_E), whose purpose is to attract the segmentation onto evident boundaries, on regions where they do exist. It is modeled as initially proposed by Caselles et al. in [15].

The second image term is the region term (E_R), modeled as in [16], which tries to create two regions (the inner and outer) of common intensity statistics, i.e. a homogeneous region with low variance, contrasting to its surroundings that have different mean value. However, in a T1-weighted brain MR image gray matter typically has lower intensity levels than white matter, but higher levels than cerebrospinal fluid (CSF). As the region term is trying to separate between two regions, it separates the white matter regions from the rest and falsely groups the lower levels containing both gray matter and CSF.

To overcome the latter, the third image term (E_G) is introduced, which acts competitively with the second term. This term is based on whole brain statistics, and classifies each voxel to CSF, gray, or white matter and it is modeled as the energy term of a region based level set applied on a slightly smoothed - to accommodate for classification errors- version of G (G_s). Thus, the purpose of the third term is to constrain the evolution to the gray matter, since HC belongs to it, and to act competitively with the second term (the second term takes into account local information, while the third term whole brain information).

The fourth term is the prior term (E_P), which constrains the evolution on the allowable space of HC. The prior term is again modeled as the energy term of a region based level-set applied on SDM (L). Region-based level sets were again chosen due to the absence of edges in this map.

The need of having those four energy terms arises from the challenging nature of HC, and each of those terms has a specific role. The HC has parts with strong edges, where the gradient term becomes useful, has other parts with weak boundaries, on which the region term would correctly guide the segmentation, and has parts with unrecognizable boundaries, where the prior term takes the lead and prevents contour leakage. The whole brain term's role is to take advantage of the more robust whole brain statistics and correct inefficiencies of the region-term, preventing it from leakage to darker voxels.

C. Tri-phase Gradient Distribution on the Boundary (3GDB)

3GDB's role is the efficient blending of the energy terms. 3GDB is a threefold weighting map that has equal dimensions with the image, i.e. $N_1 \times N_2 \times N_3 \times 3$ dimension (N_k are the dimensions of the MR image). 3GDB assigns to each voxel a specific weighting factor, defining the contribution of each update term to the contour evolution, at a voxel level. 3GDB defines the density of the gradient values on the HC boundary, thus which parts of the boundary demonstrate sufficient image information, that one should trust. Hence, on the parts with only some image information, 3GDB passes gradually the control of the contour evolution to the region terms, while on the parts with insufficient information it is passed on the prior term, in order to constrain it in the allowable space.

Fig. 3 provides an illustration of the construction of 3GDB, while the pseudo-code in Algorithm 1 provides further insights on the 3GDB extraction. More specifically, the $i-th$ training image is non-rigidly registered to the target, the transformation is applied on the label image, which is then corrected by the target's estimated gray matter image (G), as mentioned in the previous subsection. The boundary of the structure, as suggested by the corrected atlas, is extracted. The Canny edge detector [14] is applied on the target twice, with two different thresholds, once to compute only strong edges, and the other to compute both strong and weak (but existing) edges. The Canny results are binary intersected with the estimated structure's boundary, and the one is subtracted from the other. Note that in order to define the thresholds used to determine strong edges, Matlab's automatic approach was used[2]. Subsequently, to extract the weak edges, the high threshold that Matlab defined for the previous case was halved, to produce reasonable weak edges. The weak edges were subsequently defined by subtracting the strong edges from the weak and strong edges. This way we identify the parts of the structure's boundary that have significant, medium or missing gradient information, which are stored in $3GDB_j$, i.e. the first phase (A_{1j}) captures the boundary parts with strong gradients (as defined by the first Canny output), the second phase (A_{2j}) captures the boundary parts with medium strength gradients (defined by subtracting the second Canny result from the first), while the third one (A_{3j}) the parts with missing boundaries (the boundary parts where none of the Canny outputs has identified as edge). Before this operation, dilation with a cubic structuring element of size 3x3x3 was applied to accommodate for the discrepancies between the Canny's outputs (Fig. 3(c, d)) and the atlas' HC corrected boundary (Fig. 3(b)). The resulting phases (A_{1j}, A_{2j}, A_{3j}) of $3GDB_j$ are averaged over the training population, to produce three grayscale 3D images i.e. A_1, A_2, A_3, where $A_1 + A_2 + A_3 = \mathbf{1}$, which compose the 3GDB map.

D. 3GDB based update of ACM

Based on the level set analysis of [42], having an image $I \in R^3$, an evolving curve C in the image domain $\Omega \in R^3$ is defined implicitly, and is represented as the zero level set of a signed distance function $\phi : R^3 \rightarrow \Omega$

[2]http://www.mathworks.com/help/images/ref/edge.html

Data: A test image I and N atlases with their label images L_i
for *every atlas image* **do**
1. register the atlas to I ($\rightarrow$ Fig. 3(a));
2. apply the resulting transformation to L_i;
3. intersect binary $AtlasL_i$ with the gray matter of I;
4. extract the "corrected" structure's mask from step 3, and its boundary ($\rightarrow$ Fig. 3(b));
5. apply the Canny edge detector to the corrected $AtlasL_i$ to extract strong edges ($\rightarrow$ Fig. 3(c));
6. apply the Canny edge detector to the corrected $AtlasL_i$ with a lower threshold, to extract both strong and weak edges ($\rightarrow$ Fig. 3(d));
7. dilate the output of steps 5 and 6, find their intersections with the boundary of step 4, and subtract them from each other ($\rightarrow$ Fig. 3(e)), to reveal the boundary parts with strong, weak and missing edges;

end
Averaging the output of step 7 over the training set;
Result: The 3GDB (Fig. 3(f)).

Algorithm 1: Extraction of 3GDB.

(a) (b) (c) (d) (e) (f)

Fig. 3. (a) A registered atlas, (b) the registered atlas' HC boundary, after correcting it with the target's estimated gray matter, (c) the Canny edge result of (a) with a high threshold, (d) the weak gradient edges of (a), produced by binary subtracting (c) from the Canny edge result of (a) with a low threshold, (e) the high-gradient (blue) and medium-gradient (red) parts of (b), (f) the averaged over all training population three phase map 3GDB for a specific subject, after dilation, where blue depicts the regions with trustworthy gradient information, the orange to red regions depict the medium gradient information and pink the regions where prior knowledge is leading.

$$C = \{\mathbf{v} \in \Omega \mid \phi(\mathbf{v}) = 0\} \tag{2}$$

C partitions Ω into the inside to C set Ω_1, where $\phi(\mathbf{v}) < 0$, and to the outside set Ω_2, in which $\phi(\mathbf{v}) > 0$.

Given 3GDB, and its three phases A_1, A_2, and A_3, the contour update equation reads:

$$\frac{\partial \phi}{\partial t} = A_1 \circ \frac{\partial \phi_{E_E}}{\partial t} + A_2 \circ \left[\frac{\partial \phi_{E_R}}{\partial t} + \frac{\partial \phi_{E_G}}{\partial t}\right] + A_3 \circ \frac{\partial \phi_{E_P}}{\partial t} \tag{3}$$

where the operation $\circ$ denotes the Hadamard product.

E. ACM Energy Terms

1) Edge term (E_E): Based on the Geodesic Active Contours model (GAC) [15], contour evolution is regulated by the edge stopping function g, terminating it once high gradient values are detected:

$$E_E(M) = \int_{\Omega} g(\mathbf{v})|\nabla\phi(\mathbf{v})|d\mathbf{v} \tag{4}$$

The level set update term of contour ϕ based on E_E reads:

$$\frac{\partial \phi_{E_E}}{\partial t} = g|\nabla(\phi)|(\mathrm{div}\left(\frac{\nabla\phi}{|\nabla\phi|}\right) + \alpha) + \nabla g \cdot \nabla\phi \tag{5}$$

where α is the balloon force that controls shrinking or expanding of the contour. The edge stopping function used is defined as in [15] by:

$$g(|\nabla(I)|) = \frac{1}{1 + |\nabla G_\sigma * I|} \tag{6}$$

where G_σ stands for the Gaussian convolution kernel of size 3x3x3 and standard deviation 0.5. The curvature of the evolving curve $\mathrm{div}\left(\frac{\nabla\phi}{|\nabla\phi|}\right)$ acts as a regularizer for the level set by making it smooth during evolution.

2) Region term (E_R): The region-based term was modeled according to the well known Chan-Vese model [16], where for a given image $I \in \Omega$, the energy functional E_R to be minimized is formulated as:

$$E_R(M) = \lambda_1^I \int_{\Omega_1} |I(\mathbf{v}) - c_1^I|^2 d\mathbf{v} + \lambda_2^I \int_{\Omega_2} |I(\mathbf{v}) - c_2^I|^2 d\mathbf{v} \tag{7}$$

where $\mathbf{v} \in \Omega$, c_1^I, c_2^I are the average intensities of I in Ω_1 and Ω_2 respectively, while λ_1^I, λ_2^I are balancing factors between the properties of interior and exterior regions. The contour update equation of ϕ based on minimizing E_R reads:

$$\frac{\partial \phi_{E_R}}{\partial t} = \delta_\epsilon(\phi)\Big[\Big(\mu\,\mathrm{div}\Big(\frac{\nabla\phi}{|\nabla\phi|}\Big) - \nu - \lambda_1^I(I - c_1^I)^2 + \lambda_2^I(I - c_2^I)^2\Big)\Big] \tag{8}$$

where $\mu\,\mathrm{div}\left(\frac{\nabla\phi}{|\nabla\phi|}\right)$ is a regularization term added to control contour's smoothness by allowing evolution based on the curvature and thus enforcing the smoothness of the contour, ν controls the propagation speed, and $\delta_\epsilon(\phi)$ is the Dirac function.

3) Gray matter term (E_G): The purpose of this term is the inclusion of information that is based on whole brain features, indicating which voxels are likely to be included in the desired structures and which are not. To do so, a smoothed version of G (G_s), to accommodate for any small tri-segmentation error, is produced by using a Gaussian convolution kernel of size 3x3x3 and standard deviation 0.5. The same level-set contour evolved in I is also evolved in G_s, simultaneously. Having no interest on the edges of G_s (which have been smoothed), but rather on keeping the evolving contour inside the gray matter region, the Chan-Vese model is applied on G_s:

$$E_G(M) = \lambda_1^{G_s} \int_{\Omega_1} |G_s(\mathbf{v}) - c_1^{G_s}|^2 d\mathbf{v} + \lambda_2^{G_s} \int_{\Omega_2} |G_s(\mathbf{v}) - c_2^{G_s}|^2 d\mathbf{v} \tag{9}$$

where again $c_1^{G_s}$, $c_2^{G_s}$, $\lambda_1^{G_s}$ and $\lambda_2^{G_s}$ have the same nature as the ones in equation (7), while the update equation is similar with (8).

4) Prior term (E_P): In order to derive the formulation of the prior energy term, the same level-set contour evolved in I and G_s is also evolved in L, simultaneously. Since L is an image with very smooth transitions, its energy term is again modeled through the Chan-Vese model, thus becoming:

$$E_P(M) = \lambda_1^L \int_{\Omega_1} |L(\mathbf{v}) - c_1^L|^2 d\mathbf{v} + \lambda_2^L \int_{\Omega_2} |L(\mathbf{v}) - c_2^L|^2 d\mathbf{v} \tag{10}$$

where c_1^L, c_2^L, λ_1^L and λ_2^L have again the same nature as the ones in (7) and the update equation is similar with (8).

Thus, given the above, equation (3) reads:

$$\begin{aligned}\frac{\partial \phi}{\partial t} =& A_1 \circ \Big[g|\nabla(\phi)|(\mathrm{div}\Big(\frac{\nabla\phi}{|\nabla\phi|}\Big) + \alpha) + \nabla g \cdot \nabla\phi\Big] + \\ &+ \delta_\epsilon(\phi)\Big[(2A_2 + A_3) \circ \Big(\mu\,\mathrm{div}\Big(\frac{\nabla\phi}{|\nabla\phi|}\Big) - \nu\Big) - \\ &- A_2 \circ \Big(\lambda_1^I(I - c_1^I)^2 - \lambda_2^I(I - c_2^I)^2\Big) - \\ &- A_2 \circ \Big(\lambda_1^{G_s}(G_s - c_1^{G_s})^2 - \lambda_2^{G_s}(G_s - c_2^{G_s})^2\Big) - \\ &- A_3 \circ \Big(\lambda_1^L(L - c_1^L)^2 - \lambda_2^L(L - c_2^L)^2\Big)\Big]\end{aligned} \tag{11}$$

where the balancing factors of the interior and exterior regions for images G and L, for which both the interior and the exterior regions have uniform intensities, were set to be equal

($\lambda_1^{G_s} = \lambda_2^{G_s} = \lambda_1^L = \lambda_2^L = 1$). In the case of the image I, given that the foreground is quite uniform while the background is quite varying, proper segmentation (in order to ignore the background information) requires $\lambda_1^I >> \lambda_2^I$. After experimentation $\lambda_1^I = 1$, $\lambda_2^I = 0$ were set respectively. The remaining parameters were experimentally tuned and the values used were: $\mu = 0.0001$, $\nu = -0.01$, $\alpha = -1.5$.

Initialization of ϕ is provided by the Spatial Distribution Map; the most likely voxels to belong to the HC as offered by L (i.e. SDM'S voxels with maximum values), are used as seeding region.

F. Incorporating sophisticated fusion

Obviously, our methodology strongly depends on the multi-atlas concept and the fusion that involves. Fusion is used both in the extraction of the prior term and in the 3GDB calculation. In both cases, the simple and straightforward weighted averaging was used, based on the cross correlation of the target and each atlas. However, significant efforts and improvements in the field of label fusion on top of multi-atlas have been achieved in the last years. Therefore, in our methodology we have incorporated the joint label fusion strategy [50] (the authors are kindly offering their code[3]), substituting the simple weighted average both in equation (1) and in 3GDB calculation. The joint label fusion strategy is based on estimating the joint probability of two atlases making a segmentation error at a voxel, modeling this way the pairwise dependency between atlases, in an effort to minimize the total expectation of labeling error.

This way we are able to compare our contribution against fusion techniques and show that our methodology can act supplementary to sophisticated fusion concepts, leading to enhanced results.

III. Experiments

A. Datasets

Our experimentation dataset, on which we built our methodology, consists of 23 brain MR images, randomly selected from the OASIS repository [41], with the constraint to cover uniformly the whole age span. The OASIS repository offers a large number of 1.5T T1-weighted MR images of very high quality with reduced noise levels, since four scans have been collected from each individual and are averaged. The images are accompanied by demographic information and Clinical Dementia Rating (CDR), but not by ground-truth segmentations. Thus a professional radiologist has offered us his manual HC delineations, which are offered publicly[4].

To further demonstrate the efficiency of our method, we experimented with the publicly available dataset of IBSR, consisting of 18 manually segmented MRIs, and the dataset offered by the challenge conducted in the recent "Workshop on Multi-Atlas Labeling", held during MICCAI 2012 [2]. The OASIS-MICCAI dataset is another subset of the OASIS repository, which contains 15 MRIs for training purposes and 20 MRIs (coming from 15 individuals) for testing.

[3]http://www.nitrc.org/projects/picsl_malf/

[4]http://vcl.iti.gr/hippocampus-segmentation/

Method	Dice $\mu \pm \sigma$	Description
3GDB_Joint	**0.86 ± 0.04**	3GDB based ACM, joint label fusion
3GDB	0.85 ± 0.04	3GDB based ACM, weighted average fusion
ACM_Joint	0.85 ± 0.04	ACM without any local blending, joint label fusion
Multi-atlas_Joint	0.84 ± 0.04	Multi-atlas, joint label fusion
3GDB_NO	0.84 ± 0.07	3GDB based ACM, weighted average fusion, no gray matter correction
ACM	0.83 ± 0.05	ACM without any local blending, weighted average fusion
Multi-atlas	0.82 ± 0.04	Multi-atlas, weighted average fusion
ACM_NO	0.82 ± 0.08	ACM without any local blending, weighted average fusion, no gray matter correction
3GDB_NO2	0.82 ± 0.08	3GDB based ACM, weighted average fusion, no gray matter correction no gray matter term in ACM
Multi-atlas_NO	0.80 ± 0.08	Multi-atlas, weighted average fusion, no gray matter correction
Babalola et al. [8]	0.77 ± 0.07	AAM method

TABLE I
OASIS SUBSET: RESULTS USING THE MEAN DICE COEFFICIENT.

Segmentation performance is evaluated with the broadly-used Dice coefficient (D), which measures set agreement: let H be the actual volume of a structure, and $\hat{H}$ the segmentation result, then D equals:

$$D = \frac{2|\hat{H} \cap H|}{|\hat{H}| + |H|} = \frac{2 \cdot Pr \cdot Re}{Pr + Re} = \frac{2 \cdot TP}{(FP + TP) + (TP + FN)}, \qquad D \in [0, 1] \qquad (12)$$

where TP, FP and FN correspond to True Positive, False Positive and False Negative sets, respectively, while Pr and Re refer to Precision and Recall. A value of $D = 0$ indicates no overlap between the actual and estimated volume, while a value of $D = 1$ indicates perfect agreement.

B. Experiments on the OASIS dataset

To evaluate the efficiency of our methodology, the results of the very recent method from [8] (who kindly provide their implementation publicly) on this dataset are presented. The implementation of [8] includes its own training procedure with another dataset. In this work it was applied for testing, as is, in each of the OASIS MRIs, which does not allow for a straightforward comparison, but rather to be used as an indication on what sort of Dice coefficient values one should expect on it.

A series of experiments were conducted, to evaluate the behavior of our method, which are reported in Table I. Firstly, we produced segmentations without any local weighting map to blend the energy terms (abbreviated as ACM). As expected the concept of the local blending of the energy terms enhances the segmentation accuracy (Case A comparisons in Table II), by approximately 2% of Dice value in all cases with decreased standard deviation, showing that the sophisticated combination of the energy terms does enhance the segmentation performance. One-tailed paired t-tests were also performed to test

Case	Method	Compared with	p-value
A	3GDB_No	ACM_No	$1.5\text{x}10^{-4}$
	3GDB	ACM	$2.3\text{x}10^{-4}$
	3GDB_Joint	ACM_Joint	$2.0\text{x}10^{-4}$
B	Multi-atlas	Multi-atlas_No	$7.0\text{x}10^{-3}$
	3GDB	3GDB_No	$5.0\text{x}10^{-2}$
	ACM	ACM_No	$2.0\text{x}10^{-1}$
C	3GDB_No	Multi-atlas_No	$1.0\text{x}10^{-5}$
	3GDB	Multi-atlas	$1.0\text{x}10^{-4}$
	3GDB_Joint	Multi-atlas_Joint	$7.0\text{x}10^{-5}$

TABLE II
OASIS SUBSET: ONE-TAILED PAIRED T-TESTS WERE PERFORMED TO TEST FOR STATISTICALLY SIGNIFICANT IMPROVEMENTS.

the statistical significance of the offered improvements (Table II). As it can be seen, the p-values are smaller than 0.05 in all cases indicating statistical significance for the 2% performance increase.

Furthermore, we tested the contribution of the gray matter based correction on the multi-atlas (Case B comparisons), which is more than 1% in all cases, while the dispersion of the results is smaller in all cases. The corresponding one-tailed p-values are shown in Table II. Apart from correcting registration errors, we believe that in this dataset the gray matter correction tries to exclude from HC the CSF voxels contained in slightly inaccurate manual segmentations as shown in Fig. 4 (there is a 3% of HC voxels mismatch between the manual segmentations and the FSL/FAST output). In addition, the gray matter information contributes in the complete pipeline not only in performing the gray matter correction of the registered atlases, but also in the level set evolution, by introducing the gray matter term, that tries to restrict the evolution within the gray matter. To visualize the contribution of the gray matter energy term (E_G), Fig. 5 shows how it helps to exclude dark voxels, which E_R falsely introduces and E_P falsely agrees, as they do lie in highly probable regions. Hence, taking into account both global and regional information, benefits the segmentation performance. To define its overall impact on the proposed methodology, an experiment without using the gray matter neither for the gray matter correction, nor during the level-set evolution was conducted (3GDB_NO2). The resulting Dice coefficient is 0.82 whereas including the gray matter (3GDB) the corresponding Dice coefficient reaches to 0.85 (one-tailed p-value comparing 3GDB_NO2 with 3GDB is equal to 0.0047). This fact demonstrates the statistically significant positive effect of the inclusion of information regarding gray matter distribution in the proposed pipeline.

Moreover, we can observe the contribution of the 3GDB based ACM on top of the multi-atlas (Case C comparisons), which is more than 2% in all cases, and is not decreased regardless of how good the multi-atlas output is (through gray matter correction and/or joint label fusion). In addition, the improvement is statistically significant in all cases (Table II). Finally, the proposed overall contribution can be derived by comparing 3GDB vs Multi-atlas_NO, which is of 5% and half σ values, or 3GDB_Joint vs Multi-atlas_NO when the joint label fusion is included, which offers 6% ($\sigma = 0.04$ instead of $\sigma = 0.08$).

Fig. 6 shows comparison plots of 3GDB_Joint method, MA_Joint method, ACM_Joint and the AAM method of Babalola et al. [8], on every subject of the dataset, using four different metrics: the Hausdorff distance, the average undirected distance, the precision vs recall diagram and the Dice coefficient. The subjects are ranked according to ascending ground-truth HC volume, while the age of each subject and its corresponding CDR are also presented to provide an insight to the effect of volume, age and/or CDR on the segmentation accuracy. From the plot, only for the AAM method can be observed a clear bias towards performing worst in small size hippocampi. Furthermore, the agreement between the manually and automatically segmented volumes by means of the four methods was studied. Towards this, the Bland-Altman plots are also presented in the same figure, which allow the observation of the mean difference and the limits of agreement between automatically and manually segmented volumes as well as the spread of automatic-manual volume differences. The plots show that all four methods have similar width of agreement, but quite different mean values. More specifically, no significant bias between 3GDB_Joint and the manually segmented volumes can be seen, although a light tendency to overestimate small and medium sized hippocampi and to underestimate larger ones can be observed. A similar behavior can be observed for ACM_Joint method, but at the same time ACM_Joint presents a signed difference mean that is different from zero, which is not true in the case of the 3GDB_Joint method. The Bland-Altman plot comparing the ground-truth volumes and the ones obtained by Multi-atlas_Joint method shows an important underestimation bias, while for the AAM method of [8] a high overestimation bias can be noticed. These findings demonstrate that the 3GDB_Joint estimated volumes are closer to the true ones than the ones traced by the other automated methods.

Segmentation illustrations on two cases are also offered in Fig. 7, and the 3D reconstruction of the mean HC along with the errors of the four methods. Analyzing the results, it can be observed that the region around the CA1 part of the HC head is more susceptible to large errors for all methods apart from the AAM method in which case the error is spread out over the whole HC body. It is worth noticing that the

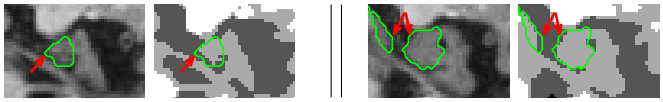


Fig. 4. Errors in the ground-truth contours in two cases (MR slice and the corresponding segmented tissues). On the images depicting the segmented tissues, white, light gray and dark gray depict white matter, gray matter and CSF respectively. The green contours depict the ground-truth boundary and the red arrows point to regions where cerebrospinal fluids were erroneously included in the hippocampal mask by the expert.

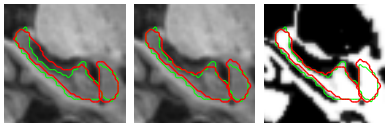
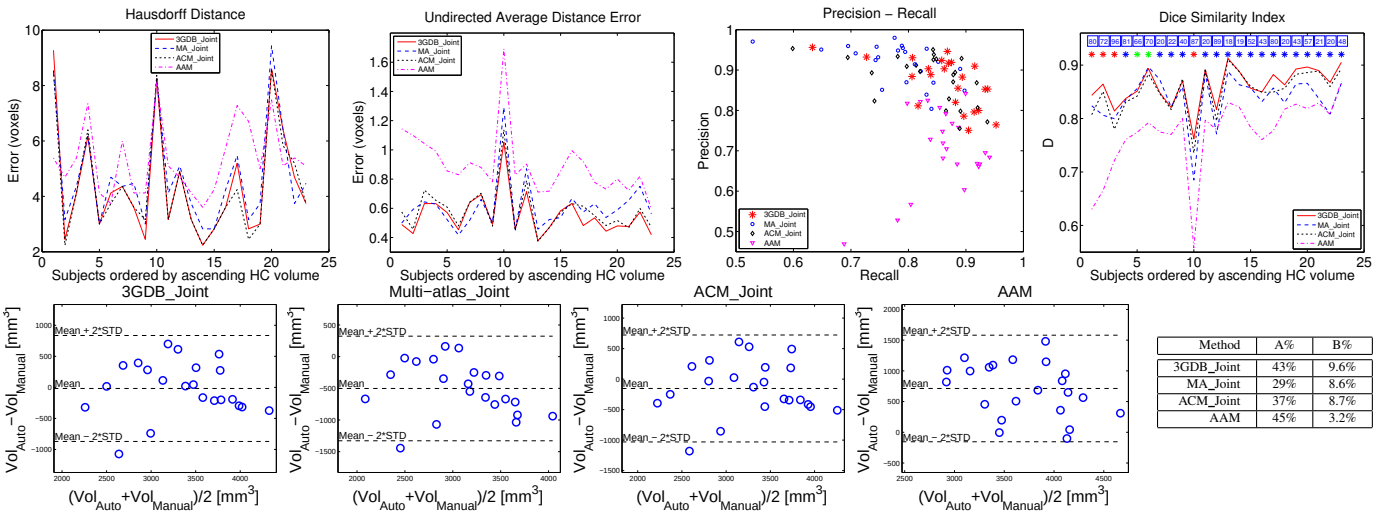


Fig. 5. Segmentation results on a sagittal slice, visualizing the benefit of including E_G. The green contour depicts the ground-truth, while the red the segmentation result excluding (left) and including (middle) E_G, respectively. On the right, the middle result is overplotted on G_s.



Method	A%	B%
3GDB_Joint	43%	9.6%
MA_Joint	29%	8.6%
ACM_Joint	37%	8.7%
AAM	45%	3.2%

Fig. 6. OASIS subset: Comparison of 3GDB_Joint, Multi-atlas_Joint, ACM_Joint and the AAM method of [8], using four metrics, and the corresponding Bland-Altman plots. The optimum position in the Precision-Recall space is the upper right corner $(1, 1)$, towards which the 3GDB_Joint asterisks have a clear tendency. On the Dice similarity index plot, the Clinical Dementia Rating (CDR) of each subject is represented by the use of colored asterisks (red asterisk stands for CDR=1, green for CDR=0.5 while blue for CDR=0). Furthermore, the age of each subject appears in a bounding box above the subject's corresponding asterisk indicating his/her CDR. Subjects are ranked by ascending ground-truth HC volume. The table on the right provides: A% the percentage of error voxels that are adjacent to the boundary, and B% the percentage of the error voxels for which the manual segmentation and FSL/FAST disagree. The figures suggest that the majority of the error voxels in all methods are not adjacent to the HC boundary, while the error that corresponds to the suspicious regions of erroneous manual segmentation is small, for all four cases.

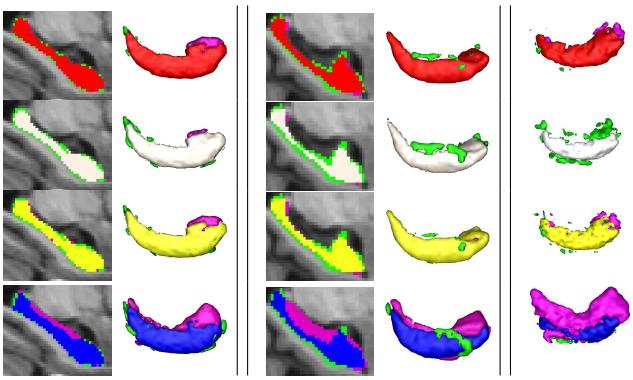


Fig. 7. OASIS subset: Visualizing results, on a sagittal slice and the corresponding 3D reconstruction of the segmented HC by means of the 3GDB_Joint (top row), the Multi-atlas_Joint method (second row), the ACM_Joint method (third row) and the AAM method of [8] (bottom row) on case 17 (left column), case 2 (middle), and the mean 3D reconstruction over the dataset (right). Pink and light green correspond to False Positives and False Negatives voxels, respectively, while red, white, yellow and blue highlight True Positives of the proposed method, the Multi-atlas_Joint method, the ACM_Joint method and the AAM method, respectively. The proposed method has an apparent efficiency on the HC-CSF borders, due to the inclusion of E_G.

Multi-atlas Joint method presents more False Negatives than 3GDB_Joint and ACM_Joint methods, which present a more similar behavior, with more concentrated error in the identified region. Furthermore, 3GDB_Joint compared to ACM_Joint method has mainly less False Negatives, but also less False Positives.

C. Comparisons on the IBSR dataset

On the IBSR dataset several other methods have published their performance during the last years, and the common metric among all is the Dice coefficient, thus a straightforward comparison becomes available, given that for our results the leave-one-out procedure was followed, using only the IBSR dataset. Those methods are primarily multi-atlas based that make use of other concepts on top of the multi-atlas. There is also a method with the very recent concept of patch-based modeling [43], and the well known FreeSurfer algorithm [26]. The results provided in Table III indicate the accuracy of the proposed method, as it produces the best published results for the task of HC segmentation.

From Table III, one can conclude the following. The enhancements of the gray matter correction in this case (comparing cases Multi-atlas vs Multi-atlas_NO) are minimal. The fact that the improvement is minimal, contrasting with the one in the OASIS subset, is due to the low imaging quality in IBSR, resulting in inaccurate FSL estimations of the gray matter. Using manually segmented gray matter masks (provided also by IBSR), instead of estimated gray matter with the use of FSL, the Multi-atlas (using the gray matter correction) and the proposed 3GDB_Joint methods reach segmentation accuracy of 0.88 and 0.89, instead of 0.83 and 0.87 when using the estimated gray matter, respectively. This result further shows that even with the most accurate gray matter estimate, there is still information that the 3GDB based ACM can exploit on top of the multi-atlas result. To further clarify the meaning of erroneous gray matter segmentation, it should be noted that 100% of the MR voxels of the manually segmented hippocampi belong to the gray matter according to IBSR ground-truth gray matter masks, whereas based on FSL estimation of gray matter, only the 78.19% belong to it. As a result, a number of HC voxels have been erroneously excluded from it during the step of gray matter correction when using the FSL-based gray matter. However, the proposed 3GDB based method still offers improvements, even by using this poor estimate. This finding suggests that utilization of gray matter correction could indeed

Method	Dice $\mu \pm \sigma$	Description
3GDB_Joint	**0.87 ± 0.02**	3GDB based ACM, joint label fusion
ACM_Joint	0.86 ± 0.02	ACM without any local blending, joint label fusion
Multi-atlas_Joint	0.85 ± 0.02	Multi-atlas, joint label fusion
3GDB	0.84 ± 0.03	3GDB based ACM, weighted average fusion
Multi-atlas	0.83 ± 0.02	Multi-atlas, weighted average fusion
Multi-atlas_NO	0.83 ± 0.03	Multi-atlas, weighted average fusion, no gray matter correction
Rousseau et al. [43]	0.83	Patch-based labelling
Lötjönen et al. [39]	0.81	Multi-atlas & intensity modeling
Sdika [44]	0.81	Multi-atlas & intensity classification
Khan et al. [35]	0.76 ± 0.03	Multi-structure registration & atlas correction
Artaech. et al. [5]	0.75	Multi-atlas with multiple combination strategies
Akselrod et al. [3]	0.69	Multi-scale segmentation with multi-atlas based prior
Fischl et al. [26]	0.75 ± 0.02	FreeSurfer

TABLE III
IBSR DATASET: COMPARISON RESULTS USING THE MEAN DICE COEFFICIENT (μ) AND THE CORRESPONDING STANDARD DEVIATION (σ) WHERE AVAILABLE.

Method	Compared with	p-value
3GDB	Multi-atlas	$5x10^{-4}$
3GDB_Joint	Multi-Atlas_Joint	$6x10^{-9}$
3GDB_Joint	ACM_Joint	$1x10^{-7}$
Multi-atlas	Multi-atlas_No	$6x10^{-4}$

TABLE IV
IBSR DATASET: STATISTICAL SIGNIFICANCE OF THE IMPROVEMENT; P-VALUES FROM PAIRED ONE-TAILED T-TESTS ARE REPORTED.

be beneficial for the task of segmentation, as long as the gray matter estimation is relatively reliable.

Furthermore, the contribution of the local weighting scheme (comparing 3GDB_Joint vs ACM_Joint) is about 1.5% and it has been proved statistically significant (Table IV).

The overall 3GDB based ACM on top of multi-atlas (comparing 3GDB vs Multi-atlas, and 3GDB_Joint vs Multi-atlas_Joint) offers statistically significant enhancements varying around 1.5-2.5% (Table IV), demonstrating a clear contribution of the proposed concept, which outperforms the previously best result of Rousseau et al. [43] about 4%.

D. Comparisons on the OASIS-MICCAI dataset

Twelve different groups submitted their works in the challenge carried out at the MICCAI 2012 workshop [2], offering 25 different implementations in total. From the published segmentation masks, we were able to extract the HC segmentation results and compute the corresponding Dice coefficients, which are reported in Table V and in Fig. 8. In order to have a fair ranking, we are using three decimal values, since the differences among the four highly ranked methods are below 1%. The challenge assessed a number of multi-atlas methods, with the differences among them primarily to be on the fusion technique they utilize. Around half of the implementations (including the 3 highly ranked) were using for the task of non-rigid registration the ANTs toolkit, thus a fair comparison with our results is available, in terms of the contributions on top of the multi-atlas.

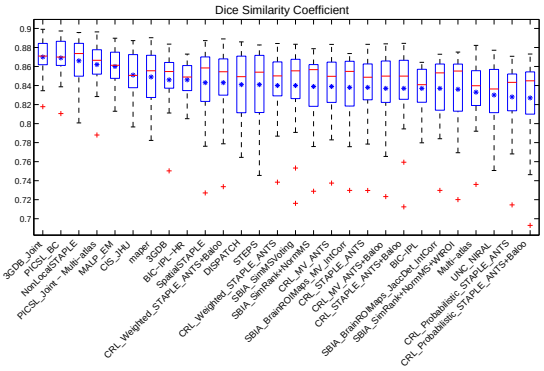


Fig. 8. OASIS-MICCAI dataset: Segmentation results in Dice based boxplot format, where the red dashes correspond to median values, while the blue asterisks to the mean values.

Method	Dice $\mu \pm \sigma$
1. *3GDB_Joint* (case iv)	**0.870 ± 0.020**
2. 'PICSL_BC'	0.869 ± 0.020
3. 'NonLocalSTAPLE'	0.866 ± 0.024
4. 'PICSL_Joint - *Multi-atlas_Joint*' (case iii)	0.862 ± 0.026
5. 'MALP_EM'	0.860 ± 0.020
7. 'CIS_JHU'	0.851 ± 0.022
8. 'maper'	0.849 ± 0.031
9. 'BIC-IPL-HR'	0.846 ± 0.018
10. *3GDB* (case ii)	0.846 ± 0.038
11. 'SpatialSTAPLE'	0.846 ± 0.018
12. 'CRL_Weighted_STAPLE_ANTS+Baloo'	0.843 ± 0.038
13. 'DISPATCH'	0.841 ± 0.036
14. 'STEPS'	0.841 ± 0.038
15. 'CRL_Weighted_STAPLE_ANTS'	0.840 ± 0.036
16. 'SBIA_SimMSVoting'	0.840 ± 0.043
17. 'SBIA_SimRank+NormMS'	0.839 ± 0.038
18. 'CRL_MV_ANTS'	0.839 ± 0.036
19. 'SBIA_BrainROIMaps_MV_IntCorr'	0.838 ± 0.037
20. 'CRL_STAPLE_ANTS'	0.838 ± 0.037
21. 'CRL_MV_ANTS+Baloo'	0.837 ± 0.041
22. 'CRL_STAPLE_ANTS+Baloo'	0.837 ± 0.042
23. 'BIC-IPL'	0.837 ± 0.022
24. 'SBIA_BrainROIMaps_JaccDet_IntCorr'	0.837 ± 0.037
25. 'SBIA_SimRank+NormMS+WtROI'	0.836 ± 0.041
26. *Multi-atlas* (case i)	0.833 ± 0.032
27. 'UNC_NIRAL'	0.830 ± 0.034
28. 'CRL_Probabilistic_STAPLE_ANTS'	0.828 ± 0.038
29. 'CRL_Probabilistic_STAPLE_ANTS+Baloo'	0.827 ± 0.044

TABLE V
OASIS-MICCAI SUBSET: RESULTS USING THE MEAN DICE COEFFICIENT.

p-values	3GDB_Joint	PICSL_BC	NonLocalSTAPLE	PICSL_Joint	MALP_EM
3GDB_Joint	—	0.8435	0.0325	0.0003	$7x10^{-8}$
PICSL_BC	0.8435	—	0.0610	0.0005	$4x10^{-7}$
NonLocalSTAPLE	0.0325	0.0610	—	0.0057	0.0090
PICSL_Joint	0.0003	0.0005	0.0057	—	0.5066
MALP_EM	$7x10^{-8}$	$4x10^{-7}$	0.0090	0.5066	—

TABLE VI
OASIS-MICCAI SUBSET: PAIRED TWO-TAILED T-TESTS WERE PERFORMED TO TEST FOR STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE FIVE TOP RATED METHODS AS PRESENTED IN TABLE V.

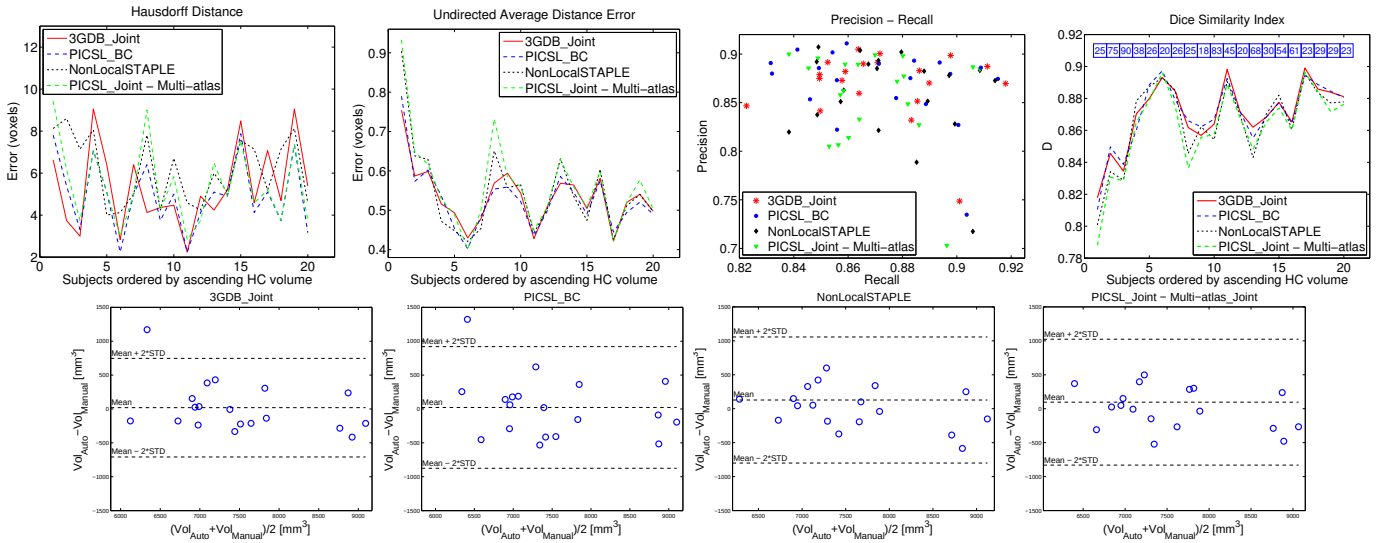


Fig. 9. OASIS-MICCAI subset: Comparison of the four top ranked methods, as presented in table V, on four metrics, and the corresponding Bland-Altman plots. Subjects are ranked by ascending ground-truth HC volume and their ages are provided in bounding boxes. Note that in the OASIS-MICCAI subset, the Clinical Dementia Rating (CDR) for all subjects is 0 or not provided by OASIS (young subjects), except for subject 16 which has CDR=0.5.

In order to compare our methodology with the published results and show our contribution, we used the same experimentation protocol with the challenge, and we have performed on this dataset four different experiments: (i) our implementation of the multi-atlas scheme based on the ANTs non-rigid registration, fusing the atlases based on the cross-correlation ANTS is providing (abbreviated as Multi-atlas), (ii) our 3GDB based ACM on top of (i) (3GDB), (iii) the ANTs based multi-atlas using the joint label fusion of [50] (Multi-atlas_Joint), and (iv) our 3GDB based ACM on top of (iii) (3GDB_Joint). It should be noted that the 'PICSL_Joint' and Multi-atlas_Joint methods are actually the same method. The only difference between them is that the first was implemented by the authors of 'PICSL_Joint', whereas the latter is our reproduction of that method, using the publicly available tools of ANTs toolkit and joint label fusion. Since both produce identical results, in all tables and figures, we refer to Multi-atlas_Joint along with 'PICSL_Joint' as PICSL_Joint- Multi-Atlas_Joint.

Comparing case (ii) to (i), and case (iv) to (iii) shows that the proposed 3GDB based ACM offers a constant and statistically significant amount of enhancement of above 1% (one-tailed p-values are equal to $3.6\text{x}10^{-8}$ and $1.9\text{x}10^{-4}$, respectively). That verifies the findings in the previous datasets, and means that the 3GDB based ACM exploits more information than the multi-atlas and the fusion scheme, regardless of how sophisticated the fusion scheme is.

Furthermore, paired two-tailed t-tests (Table VI) were used to test for significantly different means among the five top ranked methods. As demonstrated in the table, when comparing each method to its predecessor based on Dice Coefficient, the resulting outcomes did not have statistically significant differences. Significant differences were, however, observed when comparing each method to one that was placed at least two places further down the list.

In this dataset, results of the gray matter correction are not presented, as we noticed that it does not help improving the performance. In the OASIS-MICCAI dataset, the manual segmentation protocol used suggests the inclusion of white matter in HC (i.e. alveus and fimbria). Thus, inevitably, the gray matter correction cannot offer improvements in this dataset. On the contrary, the HC manual segmentation protocol in the OASIS and IBSR datasets defines HC as a homogenous gray matter structure.

Fig. 9 presents comparison plots between the four top ranked methods on every subject of the OASIS-MICCAI dataset by means of the four different metrics as in Fig. 6. All methods present a light tendency to perform better with medium and larger HC rather than with small sized ones. In the same figure, the Bland-Altman plots are provided, showing the agreement between the segmented and manually traced volumes. The plots demonstrate difference means close to zero and almost symmetric distribution around zero for 3GDB_Joint and PICSL_BC methods. Both methods therefore present no significant bias. A small overestimation bias is observed in the case of NonLocalSTAPLE and PICL_Joint methods. It is worth noticing that among all methods the spread of the automatic-manual volume differences for 3GDB_Joint is smaller than for the others methods. This fact indicates automatically segmented volumes closer to the manually segmented ones when using 3GDB_Joint method.

E. Execution time

The overall proposed pipeline requires (i) 2 hours per training image for the task of the non-rigid registration, using ANTs toolkit, (ii) almost an hour for the joint label fusion of the registered atlases, using the PICSL Multi-Atlas Segmentation Tool, (iii) 30 minutes on average to build 3GDB and, (iv) few to forty minutes for the ACM evolution, which varies due to the speed of the level set convergence, in an Intel Core i7 3.40Ghz computer, 16GB RAM. Note that cropped MR volumes were used for 3GDB construction and ACM evolution to speed up the procedure, while whole brain MR volumes were used for the tasks of non-rigid registration and

joint label fusion in order to achieve better accuracy. Thus, utilization of the proposed 3GDB concept increases slightly the computational time compared to Multi-Atlas_Joint (our contribution requires on average 45 minutes on top of the Multi-atlas_Joint, which is around 3% of the total time, when using 10 training images).

IV. Conclusions

This work proposes an ACM method on top of the multi-atlas concept. The method is based on the use of local blending of four specific energy terms for segmenting the challenging deep brain structure of HC. The novel local weighting scheme proposed intends to imitate the human expert segmentation thinking, on where and at which extent to trust image information (and which kind of it) and experience, while dealing with the multivariate nature of brain MR images. The proposed concept for modeling the varying boundary properties and subsequently choose among and blend different kind of information at a local level, can potentially be applied to other cases of similar nature, regarding the imaging properties and information, such as structures with multivariate surroundings, whose boundary demonstrates varying gradients (from very rich to even missing) at specific anatomical locations. Experiments verify the appropriateness of the specific energy terms and their local blending, since the proposed method produces better results than other state of the art methods in three datasets, one of which was used for evaluation purposes of different algorithms in a very recent segmentation challenge. Moreover, experiments verify the appropriateness of our methodology to be used on top of any multi-atlas and label fusion scheme, as it exploits structure specific information in a different way. The proposed method benefits from fusion, and acts complementary to it, as it systematically increases the segmentation accuracy.

The algorithm's most time consuming part is the non-rigid registration. Future extensions of this work would include a shift towards the patch based approaches ([23], [43]), while a very interesting idea would be to combine our concept with the bias correction concept.

Another interesting point is the fact that the method was initially designed taking into consideration manual segmentation protocols that define HC as a homogenous gray matter structure. In our OASIS subset, the manual segmentation protocol is a close variant of the protocol used in the study of [40] and considers HC as a gray matter structure. The volumetric comparisons in the IBSR dataset using the ground-truth segmentations of the gray matter and HC, have revealed that according to the used manual protocol 100% of HC is regarded as gray matter in the dataset. Currently, it is a matter of discussion and research to conclude to a common protocol that should, or not, include non-gray matter parts in the hippocampal region. The thin white matter layers of alveus and fimbria are the only non-gray matter parts that might be included, depending on the manual segmentation protocol ([33], [11]). Such a manual segmentation protocol is used in the OASIS-MICCAI subset.

Although the gray matter correction step apparently could not offer improvements in the OASIS-MICCAI subset due to the difference in the manual segmentation protocol, the proposed method has been proved efficient, even in the OASIS-MICCAI dataset and it compared favorably to the other very recent methods. This fact proves the applicability of the proposed method in both manual segmentation protocols. However, given that our motivation of utilizing the gray matter information was to exclude the CSF voxels, rather than the white matter parts, future work will focus on replacing the gray matter term with a combined gray-white matter, to accommodate for the manual protocols that include white matter in HC. Another alternative could be to identify within 3GDB the white matter borders and assign in those regions of 3GDB high weights to the prior term (which will be voting to include alveus and fimbria, as they exist also in the training set).

Concluding, given the role and importance of HC in many brain disorders, any statistical significant improvements in terms of segmentation accuracy might prove valuable, as it could lead to more reliable and more detailed biomarker identification. The proposed method by offering highly accurate HC segmentations in three different datasets (even in one with a different definition of hippocampus than the one taken into account for our method), poses a good candidate to be used in large-scale experimentation, for establishing HC volumetry as a disease biomarker.

Acknowledgment

The authors would like to thank the IBSR and the OASIS teams for providing us with their datasets. We would also like to give special thanks to Angelos Baltatzidis M.D., Radiologist for providing us with the manual segmentations of the selected OASIS MRIs. Furthermore, for the OASIS-MICCAI segmentations we would like to thank the workshop organizers, Prof. Bennett Landman and Prof. Simon Warfield, and Neuromorphometrics, Inc.

References

[1] "CAPH'08: Workshop on the Computational Anatomy and Physiology of the Hippocampus", in Medical Image Computing and Computer Assisted Intervention (MICCAI), 2008 (http://picsl.upenn.edu/caph08/).

[2] "Workshop on Multi-Atlas Labeling", in Medical Image Computing and Computer Assisted Intervention (MICCAI), 2012 (https://masi.vuse.vanderbilt.edu/workshop2012/index.php/Main_Page).

[3] A. Akselrod-Ballin, M. Galun, J.M. Gomori, A. Brandt, R. Basri, "Prior knowledge driven multiscale segmentation of brain MRI", Medical Image Computing and Computer-Assisted Intervention (MICCAI), vol. 10, 2007.

[4] P. Aljabar, R.A. Heckemann, A. Hammers, J.V. Hajnal, D. Rueckert, "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy", NeuroImage, vol. 46, pp. 726-738, 2009.

[5] X. Artaechevarria, A. Munoz-Barrutia, C. Ortiz-de-Solorzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data", IEEE Trans. on Med. Imaging, vol. 28(8), 2009.

[6] A. J. Asman, and B. A. Landman, "Multi-AtlasSegmentation using Non-Local STAPLE", MICCAI Workshop on Multi-Atlas Labeling, 2012.

[7] B. B. Avants, C. L. Epstein, M. Grossman, J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain", Medical Image Analysis, vol. 12(1), pp. 26-41, 2008.

[8] K. Babalola, T. Cootes, "Using parts and geometry models to initialise Active Appearance Models for automated segmentation of 3D medical images", IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1069-1072, 2010.

[9] 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, vol. 47(4), pp. 1435-1447, 2009.

[10] H. P. Blumberg, J. Kaufman, A. Martin, R. Whiteman, J. H. Zhang, J. C. Gore, "Amygdala and hippocampal volumes in adolescents and adults with bipolar disorder", Archives of General Psychiatry, vol. 60(12), 2003.

[11] M. Boccardi, M. Bocchetta, R. Ganzola, N. Robitaille, A. Redolfi, S. Duchesne, C. J. Jack, G.B. Frisoni, EADC-ADNI Working Group on The Harmonized Protocol for Hippocampal Volumetry and for the Alzheimer's Disease Neuroimaging Initiative, "Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation", Alzheimer's & Dementia, 2012.

[12] P. Brambilla, J. P. Hatch, J. C. Soares, "Limbic changes identified by imaging in bipolar patients", Current Psychiatry Reports, vol. 10(6), 2008.

[13] X. Bresson, P. Vandergheynst, J. P. Thiran , "A Variational Model for Object Segmentation Using Boundary Information and Shape Prior Driven by the Mumford-Shah Functional", Int. J. of Computer Vision, vol. 68(2), pp. 145-162, 2006.

[14] J. Canny, "A Computational Approach to Edge Detection", IEEE Trans. Pattern Anal. Mach. Intell., vol. 8(6), pp. 679-698, 1986.

[15] V. Caselles, R. Kimmel, G. Sapiro, "Geodesic active contours", Int. J. of Computer Vision, vol. 22(1), pp. 61-79, 1997.

[16] T. Chan and L. Vese, "Active contours without edges", IEEE Transactions on Image Processing, vol. 10, pp. 266-277, 2001.

[17] M. Chupin, A. R. Mukuna-Bantumbakulu, D. Hasboun, E. Bardinet, S. Baillet, S. Kinkingnehun, L. Lemieux, B. Dubois, and L. Camero,"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, vol. 34(3), pp. 996-1019, 2007.

[18] M. Chupin, E. Gérardin, R. Cuingnet, C. Boutet, L. Lemieux, S. Lehéricy, H. Benali, L. Garnero, and O. Colliot, "Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI", Hippocampus, 19(6), pp. 579-587, 2009.

[19] 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, vol. 52(4), 2010.

[20] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active shape models-Their training and applications", Computer Vision and Image Understanding, vol. 61, pp. 38-59, 1995.

[21] T. F Cootes, D. G. Edward, C. J Taylor, "Active appearance model", Proc. of European Conference on Computing and Visualization, 1998.

[22] T. F. Cootes, D. J. Edwards, and C. J. Taylor, "Active Appearance Models", IEEE Trans. Pattern Anal. Mach. Intell., vol. 23(6), 2001.

[23] 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, vol. 54(2), pp. 940-954, 2011.

[24] N. A. DeCarolis, A. J. Eisch, "Hippocampal neurogenesis as a target for the treatment of mental illness: a critical evaluation", Neuropharmacology, vol. 58(6), pp. 884-93, 2010.

[25] G. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models", in Proc. Third IEEE Intern. Conf. on Automatic Face and Gesture Recognition, pp. 300-305, 1998.

[26] B. Fischl, D. H Salat, E. Busa, M. Albert, M. Dieterich, et al., "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain", Neuron, vol. 33(3), pp. 341-355, 2002.

[27] X. Gao, Y. Su, X. Li, D. Tao, "A Review of Active Appearance Models", IEEE Trans. on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 40(2), pp. 145-148, 2010.

[28] 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, vol. 33(1), 2006.

[29] T. Heimann and H. P. Meinzer,"Statistical shape models for 3D medical image segmentation: A review", Med. Im. Anal., vol. 13(4), 2009.

[30] S. Hu, P. Coupé, J. Pruessner, D. L. Collins, "Validation of appearance-model based segmentation with patch-based refinement on medial temporal lobe structures", MICCAI Workshop on Multi-Atlas Labeling and Statistical Fusion, 2011.

[31] S. Hu, P. Coupé, J. Pruessner and D.L. Collins, "Appearance-based Modeling for Segmentation of Hippocampus and Amygdala using Multi-contrast MR Imaging", NeuroImage, vol. 58(2), pp. 549-559, 2011.

[32] S. Hu and D. L. Collins, "Joint level-set shape modeling and appearance modeling for brain structure segmentation", NeuroImage, vol. 36, 2007.

[33] C. Konrad, T. Ukas, C. Nebel , V. Arolt, A. W. Toga, K. L. Narr, "Defining the human hippocampus in cerebral magnetic resonance images - An overview of current segmentation protocols", NeuroImage 47, 2009.

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

[35] A. R. Khan, M. K. Chung, M. F. Beg, "Robust atlas-based brain segmentation using multi-structure confidence-weighted registration", Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), vol. 12, pp. 549 -557, 2009.

[36] C. Langan, C. McDonald, "Neurobiological trait abnormalities in bipolar disorder", Molecular Psychiatry, vol. 14(9), pp. 833-846, 2009.

[37] J. M. Leventon, E. Grimson, and O. Faugeras, "Statistical shape influence in geodesic active contours", IEEE Conf. on Computer Vision Pattern Recognition, vol. 1, 2000.

[38] F. Van der Lijn, T. den Heijer, M. M. B. Breteler, W. J. Niessen, "Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts", NeuroImage, vol. 43(4), 2008.

[39] J. M. Lötjönen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, D. Rueckert, "Fast and robust multi-atlas segmentation of brain magnetic resonance images", Neuroimage, vol. 49(3), 2010.

[40] K. L Narr, P. M. Thompson, P. Szeszko, et al., "Regional specificity of hippocampal volume reductions in first-episode schizophrenia", Neuroimage 21, pp. 1563-75, 2004.

[41] D. S. Marcus, T. H. Wang, J. Parker, J. G Csernansky, J. C. Morris, R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults", J of Cognitive Neuroscience, vol. 19, 2007.

[42] S. Osher, R. Fedkiw, "Level Set Methods and Dynamic Implicit Surfaces", Springer Verlag, 2002.

[43] F. Rousseau, P. A. Habas, C. Studholme, "A supervised patch-based approach for human brain labeling", IEEE Trans. Med. Imag., vol. 30(10), pp. 1852-1862, 2011.

[44] M. Sdika, "Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote", Medical Image Analysis, vol. 14(2), pp. 219-226, 2010.

[45] S. M. Smith, " Fast robust automated brain extraction", Human Brain Mapping, vol. 17(3), pp. 143-155, 2002.

[46] M. B. Stegmann, R. Larsen, "Multi-band Modelling of Appearance", Image and Vision Computing, vol. 21(1), pp. 61-67, 2003.

[47] A. Sumich, X. A. Chitnis, D. G. Fannon, S. O'Ceallaigh, V. C. Doku, A. Falrowicz, "Temporal lobe abnormalities in first-episode psychosis", Am J Psychiatry, vol. 159(7), pp. 1232-1235, 2002.

[48] Z. Tu, K. L. Narr, P. Dollar, I. Dinov, P. M. Thompson, A. W. Toga, "Brain Anatomical Structure Segmentation by Hybrid Discriminative/Generative Models", IEEE Trans. Med. Imag., vol. 27(4), 2008.

[49] D. Velakulis, S. J. Wood, M. T. H. Wong, et al., "Hippocampal and Amygdala volume according to Psychosis stage and diagnosis", Archives of General Psychiatry, vol. 63(2), pp. 139-49, 2006.

[50] H. Wang, J. W. Suh, S. R. Das, J. Pluta, C. Craige and P. A. Jushkevich, "Multi-Atlas Segmentation with Joint Label Fusion", IEEE Trans. Pattern Anal. Mach. Intell., vol. 35(3), 2013.

[51] H. Wang, B. Avants, and P. A. Jushkevich, "Grand Challenge on Multi-Atlas Segmentation: A Combined Joint Label Fusion and Corrective Learning Approach", MICCAI Workshop on Multi-Atlas Labeling, 2012.

[52] J. Yang, L. H. Staib and J. S. Duncan, "Neighbor-Constrained Segmentation with Level Set Based 3D Deformable Models", Trans. Med. Imag, vol. 23(8), 2004.

[53] D. Zarpalas, A. Zafeiropoulos, P. Daras, N. Maglaveras, M. G. Strintzis, "Brain Structures Segmentation using Optimum Global and Local Weights on Mixing Active Contours and Neighboring Constraints", Int. Symp. on Applied Sciences in Biomedical and Communication Technologies (ISABEL), 2011.

[54] D. Zarpalas, P. Gkontra, P. Daras, and N. Maglaveras, "Hippocampus Segmentation through Gradient based Reliability Maps for Local Blending of ACM Energy Terms", IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro, 2013.

[55] Y. Zhang, M. Brady, S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectationmaximization algorithm", Trans. Med. Imag., vol. 20(1), 2001.