MULTI-CLASS BRAIN SEGMENTATION USING ATLAS PROPAGATION AND EM-BASED REFINEMENT

Christian Ledig[⋆]*, Robin Wolz*[⋆]*, Paul Aljabar*[⋆]*, Jyrki Lötjönen*[†]*, Rolf A. Heckemann*[‡]*,*
Alexander Hammers[‡]*, Daniel Rueckert*[⋆]

[⋆] Department of Computing, Imperial College London, London, UK
[†] Knowledge Intensive Services, VTT Technical Research Centre of Finland, Tampere, Finland
[‡] The Neurodis Foundation, Lyon, France

ABSTRACT

In recent years, multi-atlas segmentation has emerged as one of the most accurate techniques for the segmentation of brain magnetic resonance (MR) images, especially when combined with intensity-based refinement techniques such as graph-cut or expectation-maximization (EM) optimization. However, most of the work so far has focused on intensity-based refinement strategies for individual anatomical structures such as the hippocampus. In this work we extend a previously proposed framework for labeling whole brain scans by incorporating a global and stationary Markov random field that ensures the consistency of the neighbourhood relations between structures with an a-priori defined model. In particular we improve the segmentation result of a locally weighted multi-atlas fusion method for 41 different structures simultaneously by applying a subsequent EM optimization step. We evaluate the proposed approach on 30 manually annotated brain MR images and observe an improvement of label overlaps to a manual reference by up to 6% . We also achieved a considerably improved group separation when the proposed segmentation framework is applied to a volumetric analysis of 404 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort.

Index Terms— segmentation, atlas-propagation, EM optimization, Markov random field, simultaneous refinement

1. INTRODUCTION

The accurate labeling of anatomical structures in MR brain scans is a fundamental task in neuroimaging. In order to use measurement of regional volumes to assist in clinical decision-making, precise segmentation is crucial. For dementia disease in particular, correlations between region volumes and disease progress have been shown. For example, hippocampal atrophy is a good indicator and predictor for Alzheimer's disease (AD). Manual segmentation of whole brain images is not scalable, since it requires substantial time, effort and expertise. There is thus a pressing need for automated processes that minimize the necessity of user interaction.

Good segmentation results are obtained by multi-atlas label propagation [1]. This approach is based on so called *atlases* consisting of brain MR images and corresponding manual labels that have been prepared using detailed manual segmentation protocols [2]. Finding the transformations between a set of atlas images and an unseen image, allows the mapping of each atlas segmentation into the space of the unseen image. Every transformed atlas provides an *opinion* in form of a vote on the label of each voxel in the target image. Different strategies have been proposed to fuse the individual votes into a consensus segmentation. Examples range from simple majority voting to more sophisticated methods such as neighbourhood-dependent local weighting strategies [3].
A contrasting, yet equally successful approach to image segmentation is based on a probabilistic intensity model which is solved with an EM optimization method. This technique is widely used and can be extended to also model the non-uniform bias field that is usually present in MR images [4].

In this paper, we extend our previous work [5] that combines both strategies by employing the segmentation obtained through multi-atlas fusion as spatial priors for an intensity model solved by EM optimization. We add a global and stationary Markov random field (MRF) that enforces a consistent segmentation by incorporating prior knowledge on the neighbouring graph of different brain structures into the segmentation process. Additionally, we employ a local weighting scheme to define subject-specific probabilistic atlases and evaluate the method on a whole-brain atlas defining 41 structures. Using a probabilistic atlas that is specifically generated for a given target image, contrasts our method from multi-structure segmentation techniques like the one described by Fischl et al [6]. We focus on cortical grey matter structures since large structural intersubject variation and strong partial volume (PV) effects in these regions tend to hamper non-rigid registration approaches on which atlas propagation techniques rely.

2. METHOD

2.1. Intensity Model and ML Approach

For the sake of consistency and readability we follow the notations used by van Leemput et al. [4]. We consider the following model of an MR image: Let the n voxels of an image be indexed by $i = 1, \ldots, n$. Given an intensity $y_i \in \mathbb{R}^m$ an image is well defined by $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$. In the following we consider the single channel case ($m = 1$). For a given image $\mathbf{y}$ the corresponding segmentation is given as $\mathbf{z} = \{\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_n}\}$, where $\mathbf{z_i} = \mathbf{e_k}$ with $1 \leq k \leq K$ describes the class k to which voxel i belongs and K is the number of tissue classes. $\mathbf{e_k}$ denotes a unit vector having the k^{th} component equal to 1 while all others are 0.

The challenge is to estimate the underlying but unknown segmentation $\mathbf{z}$ by means of the observed intensities $\mathbf{y}$. We follow the common assumption [4, 7] that the observed log-transformed intensities of voxels belonging to a certain class k are normally distributed

*This work is partially funded under the 7th Framework Programme by the European Commision (http://cordis.europa.eu/ist/)

with mean μ_k and standard deviation σ_k. This gives the overall model parameters as $\mathbf{\Phi} = \{(\mu_1, \sigma_1), (\mu_2, \sigma_2), \ldots, (\mu_K, \sigma_K)\}$.

Assuming that the parameters $\mathbf{\Phi}$ are known the probability of observing an intensity y_i at voxel i can be computed as:

$$f(y_i|\mathbf{\Phi}) = \sum_k f(y_i|\mathbf{z_i} = \mathbf{e_k}, \mathbf{\Phi}) f(\mathbf{z_i} = \mathbf{e_k}) \tag{1}$$

Here $f(\mathbf{z_i} = \mathbf{e_k})$ is the probability of voxel i having tissue class k and $f(y_i|\mathbf{z_i} = \mathbf{e_k}, \mathbf{\Phi})$ is a Gaussian distribution with parameters (μ_k, σ_k). At this point we assume statistical independence between all voxels. The overall probability of observing an image $\mathbf{y}$ with known parameters $\mathbf{\Phi}$ can then be written as $f(\mathbf{y}|\mathbf{\Phi}) = \prod_i f(y_i|\mathbf{\Phi})$. We can now apply the EM algorithm by interleaving the expectation of the class probabilities $p_{ik}^{(m+1)}$, while assuming that the model parameters $\mathbf{\Phi}$ are fixed, and the maximization of $f(\mathbf{y}|\mathbf{\Phi})$. m denotes the iteration number. In order to account for topological and spatial apriori knowledge this model can be further modified by employing Markov random fields (MRF). This is done by expanding $f(\mathbf{z_i} = \mathbf{e_k})$ to

$$f(\mathbf{z_i} = \mathbf{e_k}|p_{\mathcal{N}_i}^{(m)} \Phi_z^{(m)}) = \frac{\pi_{ik} e^{-U_{MRF}(\mathbf{e_k}|p_{\mathcal{N}_i}^{(m)}, \Phi_z^{(m)})}}{\sum_{j=1}^{K} \pi_{ik} e^{-U_{MRF}(\mathbf{e_k}|p_{\mathcal{N}_i}^{(m)}, \Phi_z^{(m)})}} \tag{2}$$

where π_{ik} is an anatomical probabilistic prior for class k at voxel i and $\mathcal{N}_i$ is the set of first-order neighbours of voxel i.

We follow the approach of global and stationary MRF taken in [7] and employ the energy function

$$U_{MRF}(\mathbf{e_k}|p_{\mathcal{N}_i}^{(m)}, \Phi_z^{(m)}) = \tag{3}$$

$$\sum_{j=1}^{K} G_{kj} \Big(\sum_{l \in \mathcal{N}_i^x} s_x p_{lj} + \sum_{l \in \mathcal{N}_i^y} s_y p_{lj} + \sum_{l \in \mathcal{N}_i^z} s_z p_{lj} \Big) \tag{4}$$

where G denotes a $K \times K$ matrix defining the connectivity between class k and j and $s = \{\frac{1}{d_x}, \frac{1}{d_y}, \frac{1}{d_z}\}$ accounts for the anisotropic voxel spacing in world coordinates.

2.2. Multi Atlas Propagation

To incorporate anatomical priors into the model we derive a digital brain atlas for the unseen image $\mathbf{y^u}$ that is to be classified. After transforming N atlases to the coordinate system of image $\mathbf{y^u}$, we create a probabilistic atlas by employing a locally weighted multi-atlas fusion strategy [3]. We denote the transformed atlases as A^n with $n = 1, \ldots, N$. After a rigid and affine alignment step, a non-rigid registration based on free-form deformations (FFD) [8] is applied to align all atlas images with the target. Normalized mutual information (NMI) is used as a similarity measure and non-rigid registration is carried out in a multi-resolution fashion using B-spline control point grids with spacings of 20mm, 10mm, 5mm and 2.5mm. A probabilistic atlas is then created by locally weighting each propagated atlas according to its local similarity to the target image.

The local similarity measure we employ is the sum of squared differences (SSD) in a cubic r-neighbourhood, $\mathcal{N}_{i,r}$, around each voxel since it outperforms most other fusion rules according to [3]. We computed this local similarity on rescaled, intensity-normalized images. For K classes and N atlases we compute the probabilistic atlas at voxel i for class k as

$$\pi_{ik} = \frac{\sum_{n=1}^{N} w_{in} \gamma_{ik}^n}{\sum_{k=1}^{K} \sum_{n=1}^{N} w_{in} \gamma_{ik}^n} \tag{5}$$

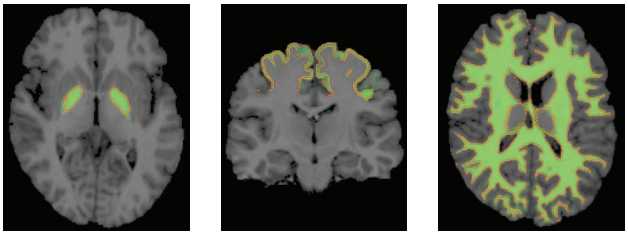


Fig. 1. Probabilistic priors (green) for the pallidum (left), for the precentral gyrus (a cortical class, middle) and white matter (right) with overlayed manual segmentation (red).

with

$$w_{in} = \Big(\sum_{j \in \mathcal{N}_{i,r}} (y_j^u - A_j^n)^2 \Big)^{-1}, \quad \gamma_{ik}^n = \begin{cases} 1, & \text{if } \mathbf{z_i} = \mathbf{e_k} \text{ in } A^n \\ 0, & \text{else} \end{cases} \tag{6}$$

Examples of the probabilistic atlas and manual segmentations are shown in Figure 1.

3. VALIDATION

3.1. Manually Segmented Data

We evaluated the approach presented in Section 2 on 30 manually segmented T_1-weighted MRI scans. Each brain image had been segmented by experts into 83 different regions [2], distinguishing equivalent areas in the right and left hemisphere. In order to employ these segmentations for grey matter segmentation we conducted two major pre-processing steps:

a) *Fusion of corresponding structures:* On the assumption that paired structures that appear in both the left and right hemispheres (e.g. the thalamus) have the same intensity characteristics, we applied a common label to each pair. This step reduced the initial segmentation from 83 to 43 classes, since brainstem, third ventricle and corpus callosum are unpaired structures.

b) *Grey matter segmentation:* The manual segmentation is defined by protocols based on anatomy rather than intensity. To adapt the atlases to our framework, most anatomical regions were subdivided into their white matter (WM), grey matter (GM) and cerebrospinal fluid (CSF) portions using FSL FAST 4.1.8 tissue classification [9] on robustly intensity-scaled versions of the atlas T_1 MR images. Exceptions were several regions expected to contain mainly one type of class (corpus callosum, ventricles, hippocampus (Hi), amygdala (Am), caudate nucleus (CauNuc), nucleus accumbens (NucAcc), substantia nigra (SubNig)), or to contain a mix of GM/WM, where reclassifying would not properly represent the actual structure (putamen (Put), thalamus (Tha), and pallidum (Pal)). We then merge all WM structures together with the corpus callosum into a general WM label and all CSF structures together with the ventricles into a CSF label. We thus obtain 8 subcortical GM structures, 29 cortical GM structures and the GM fractions of the brainstem and of the cerebellum, which with CSF and WM results in a total of 41 labels.

3.2. Evaluation Strategy and Parameters

We evaluated the EM-based refinement against multi-atlas propagation with majority vote fusion as well as against multi-atlas propa-

	MVF	+EM	LWF	+EM
SI hippocampus	82.7	81.6	83.6	82.3
SI 41 classes	83.6	86.6	84.1	**86.9**
SI 39 all GM classes	78.2	82.0	78.9	**82.5**
SI 29 cort. GM classes	75.0	79.4	75.7	**79.9**
WTC 41 classes	59.7	64.0	60.6	**64.7**
WTC 39 all GM classes	58.9	63.2	59.8	**63.9**
WTC 29 cort. GM classes	56.1	61.5	56.8	**62.1**

Table 1. Similarity Indices (SI) and weighted Tanimoto Coefficients (WTC) for all 41, all 39 GM and all 29 cortical GM classes obtained with majority vote fusion (MVF), MVF&EM, local weighted fusion (LWF), LWF&EM. (bold = significant at $p = 0.05$)

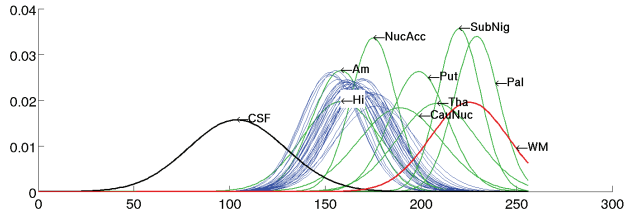


Fig. 2. Gaussian intensity distribution of all 41 classes from the manual segmentation. CSF (black), WM (red), deep GM structures (green), cortical GM, brainstem and cerebellum GM fractions (blue).

gation with a locally weighted fusion as presented in Section 2.2. To identify neighbouring tissue classes for the implementation of the MRF, we counted the labels of adjacent voxels in the manual segmentations. After thresholding we obtained the 41×41 matrix G (see Eq. 4) with entry (i, j) defined as:

$$G(i,j) = \begin{cases} 0, & \text{if } i = j \\ 0.15, & \text{if structures } i \text{ and } j \text{ share a boundary} \\ 0.75, & \text{if structures } i \text{ and } j \text{ are distant} \end{cases} \quad (7)$$

We run an evaluation for each of the 30 images using a leave-one-out strategy. The EM algorithm terminates if the change of the loglikelihood of subsequent iterations is less than 0.5%. For the fusion of the 29 propagated atlases in each test we used $r = 5$ to compute the local similarity (see Eq. 6).

3.3. Overlap Metrics

We computed the Similarity Index (SI, alternatively known as Dice coefficient) as well as the weighted Tanimoto Coefficient (WTC) [10], which takes the volume differences between labels into account. As shown in Table 1, we observed significantly increased overlaps in cortical GM structures compared with multi-atlas propagation, regardless of which fusion strategy was employed. The largest overlap improvements were seen in cortical GM classes, with similarity indices increased by 2-6% for each individual cortical GM structure. This aligns with our expectations, since the intensity histograms of CSF, cortical GM and WM are well separated [6]. In subcortical GM structures such as the thalamus or the pallidum, the EM-based refinement step did not lead to consistent improvements. There are a number of possible reasons for this: Firstly, as shown in Figure 2 and also [6], the separation of WM from subcortical GM is generally less clear than the separation of WM from cortical GM.

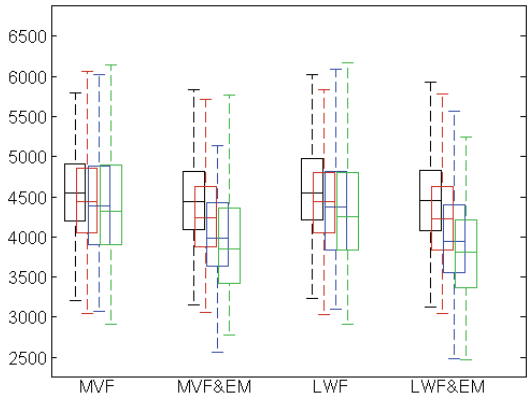


Fig. 3. Boxplot of hippocampal volumes of, from left to right, healthy control (black), MCI non-converter (red), MCI converter (blue) and AD (green) subjects obtained with each method.

Another problem for the EM algorithm is that boundaries between neighbouring subcortical structures auch as the hippocampus and the amygdala can be hard to refine since there is no substantial difference in intensity distribution [6]. Finally, the fact that in subcortical regions multi-atlas propagation approaches the accuracy of interrater reliability (0.85 and above) makes the objective measurement of further improvements difficult. Our experiments corroborate the results presented in [3], showing that a local weighted fusion strategy can outperform simple majority voting.

3.4. Intra-class Variance

Experiments revealed that the intensity variance within most labels is significantly higher in the segmentation obtained through multi-atlas propagation compared to the manual segmentation. The proposed refinement step reduced this intra-class variance for all labels when compared to multi-atlas propagation, and for a majority of labels when compared to manual segmentation. A lower intra-class variance does not, however, imply a better segmentation, as can be seen in comparison to the manual segmentation. Deep GM structures in particular, such as the putamen or thalamus, may already have a rather high intra-class variance. Nevertheless a better separation between structures in terms of intensity might be beneficial in some applications.

3.5. Group separation using ADNI data

To further investigate potential benefits of an improved class separation, we evaluated our method on 797 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI, adni.loni.ucla.edu) database [11]. The cohort consisted of 220 healthy controls, 393 subjects with mild cognitive impairment (MCI) and 184 AD patients. The MCI group is further divided into 229 non-converters and 164 converters respectively. After registering [1] the 30 atlases (compare Section 3.1) to each of the 797 subjects, we applied our method as introduced in Section 2. For each dataset we then computed both the volume of the hippocampus and the amygdala. Computing the hippocampal volume of all subjects revealed an improved separation of the healthy control, MCI non-converter, MCI converter and AD groups for methods using the EM refinement. The differences between the resulting volumes of fusion techniques and EM refined methods are

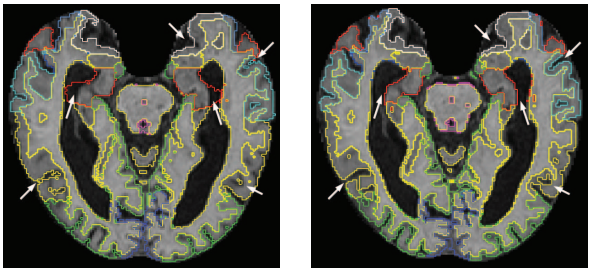


Fig. 4. Segmentation results obtained on a subject of the ADNI cohort with locally weighted multi-atlas fusion without (left) and with (right) EM refinement.

	MVF	+EM	LWF	+EM
hippocampus	61%	69%	63%	**72%**
amygdala	72%	74%	72%	**75%**

Table 2. Classification accuracy of 404 ADNI subjects (184 AD patients, 220 healthy controls) based on the volumes of the hippocampus and amygdala.

statistically significant on a 5% significance level ($p < 10^{-5}$) for all groups. A boxplot illustrating these results is provided in Figure 3. It also shows the improved separation of MCI converters from MCI non-converters if EM refinement is used.

Additionally we evaluated our method by classifying 184 AD patients and 220 healthy controls, where we used a classifier based on simple thresholding of the volumes of the hippocampus or the amygdala. We trained the classifier on 85 % of the 404 subjects and tested with the remaining 15%. The results in Table 2 were obtained by cross-validation (n=1000) and are in line with results achieved by segmenting the ADNI database with the same atlas in our LEAP framework [12]. Apart from increasing classification rates considerably, we confirmed the potential of morphometry of the amygdala for AD classification. The rather low value for the hippocampal region with multi-atlas fusion may be caused by failures of the non-rigid registration step on subjects with severe atrophy of the mesial temporal lobe. Since we are not using any atlas selection or intermediate templates, this is not surprising. However it is noteworthy that the refinement step is able to partly compensate these registration problems for this particular application. Visual results shown in Figure 4 reveal improved separation in the cortical area of the EM-refined labeling and a corrected segmentation in the hippocampal area.

4. CONCLUSION

The refinement of a segmentation obtained through multi-atlas propagation by employing EM optimization techniques can lead to major improvements in segmentation quality. With the proposed approach we are able to refine 41 anatomical structures simultaneously. The method provides consistent improvements in overlap metrics for all cortical GM classes. In cortical regions, the similarity index increased by an average of 4.2% compared to the results obtained through locally weighted multi-atlas fusion. Apart from overlap comparisons we also carried out statistical evaluations, revealing a significant and consistent reduction in the intra-class variance for all structures. We showed that our method can improve classification into diagnostic groups based on regional grey matter volumetry. Since our method allows us to incorporate the volumes of 41 different structures instead of one single structure there is also the potential to build more complex classifiers. In future work it will be interesting to investigate how segmentation results of other techniques, for example methods based on a step-wise registration procedure (LEAP, [12]), can be refined and how the additional refinement of the priors could lead to more sophisticated segmentation methods.

5. REFERENCES

[1] R. A. Heckemann, S. Keihaninejad, P. Aljabar, K. R. Gray, C. Nielsen, D. Rueckert, J. V. Hajnal, A. Hammers, and The AD Neuroimaging Initiative, "Automatic morphometry in Alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 56, no. 4, pp. 2024–2037, Mar. 2011.

[2] A. Hammers, R. Allom, M. Koepp, S. L. Free, R. Myers, L. Lemieux, T. N. Mitchell, D. J. Brooks, and J. S. Duncan, "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe," *Human Brain Mapping*, vol. 19, no. 4, pp. 224–247, 2003.

[3] X. Artaechevarria, A. Munoz Barrutia, and C. Ortiz de Solorzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE TMI*, vol. 28, no. 8, pp. 1266–1277, Aug. 2009.

[4] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE TMI*, vol. 18, no. 10, pp. 897 –908, 1999.

[5] J. MP. Lötjönen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert, "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *Neuroimage*, vol. 49, no. 3, pp. 2352 – 2365, 2010.

[6] B. Fischl, D.H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A.M. Dale, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, Jan. 2002.

[7] M. J. Cardoso, M. J. Clarkson, G. R. Ridgway, M. Modat, N. C. Fox, and S. Ourselin, "Load: A locally adaptive cortical segmentation algorithm," *Neuroimage*, vol. 56, no. 3, pp. 1386 – 1397, 2011.

[8] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE TMI*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[9] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm," *IEEE TMI*, vol. 20, no. 1, pp. 45–57, 2001.

[10] W.R. Crum, O. Camara, and D.L.G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE TMI*, vol. 25, no. 11, pp. 1451 –1461, nov. 2006.

[11] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C. Jack, W. Jagust, J.Q. Trojanowski, A.W. Toga, and L. Beckett, "The Alzheimer's Disease Neuroimaging Initiative," *Neuroimaging Clinics of North America*, vol. 15, no. 4, pp. 869 – 877, 2005.

[12] R. Wolz and D. Rueckert, "Segmentation of the ADNI database into 83 anatomical regions using LEAP," Toronto, Canada, 2011, MICCAI 2011 Workshop on MALSF.