NeuroImage: Clinical 12 (2016) 570–581

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Direct estimation of patient attributes from anatomical MRI based on multi-atlas voting

Dan Wu[a], Can Ceritoglu[b], Michael I. Miller[b,c,d], Susumu Mori[a,e,*]

[a] Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[b] Center for Imaging Science, Johns Hopkins University, Baltimore, MD, USA
[c] Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, USA
[d] Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[e] F.M. Kirby Research Center for Functional Brain Imaging, Kennedy Krieger Institute, Baltimore, MD, USA

ARTICLE INFO

Article history:
Received 25 July 2016
Received in revised form 7 September 2016
Accepted 8 September 2016
Available online 14 September 2016

Keywords:
Multi-atlas voting
Context-based image retrieval
Diagnostic estimation
Atlas-weighting
Alzheimer's disease

ABSTRACT

MRI brain atlases are widely used for automated image segmentation, and in particular, recent developments in multi-atlas techniques have shown highly accurate segmentation results. In this study, we extended the role of the atlas library from mere anatomical reference to a comprehensive knowledge database with various patient attributes, such as demographic, functional, and diagnostic information. In addition to using the selected (heavily-weighted) atlases to achieve high segmentation accuracy, we tested whether the non-anatomical attributes of the selected atlases could be used to estimate patient attributes. This can be considered a context-based image retrieval (CBIR) approach, embedded in the multi-atlas framework. We first developed an image similarity measurement to weigh the atlases on a structure-by-structure basis, and then, the attributes of the multiple atlases were weighted to estimate the patient attributes. We tested this concept first by estimating age in a normal population; we then performed functional and diagnostic estimations in Alzheimer's disease patients. The accuracy of the estimated patient attributes was measured against the actual clinical data, and the performance was compared to conventional volumetric analysis. The proposed CBIR framework by multi-atlas voting would be the first step toward a knowledge-based support system for quantitative radiological image reading and diagnosis.

1. Introduction

In this paper, we developed and tested a multi-atlas voting method that can capture brain anatomical features that are associated with certain functional deficits or diagnostic categories of dementia patients, based on existing knowledge database (atlases), and applied this knowledge to individual patients to support clinical image reading. Anatomical MRI is an indispensable tool to diagnose various brain diseases. Three types of MRI methods—T1-weighted, T2-weighted, and FLAIR—have been the most widely used clinically. Based on specific features that appear in these images, radiologists estimate the likely causes of the features and arrive at the best medical judgment. There are three types of critical information radiologists extract from the images: the type, degree, and location of the features. These features are then compared to the radiologist's knowledge about the range of normal appearance at a given age of the patient. If considered abnormal, the type, degree, and location of the abnormality are documented in a radiological report. Radiologists often go one step further by performing a similarity search within their knowledge base of various diseases and provide potential diagnoses. In the field of computer vision, this is a type of context-based image retrieval (CBIR) (Hsu et al., 2012; Muller et al., 2004; Smeulders et al., 2000; Tang et al., 1999). Namely, there is a knowledge database that contains images and associated text-based attributes, such as demographic, functional, and diagnostic information. When an image of a new patient is provided, along with his/her demographic and clinical information, past cases with similar features are extracted, together with the desired diagnostic information. The long-term goal of this study is to develop a new image analysis tool that will emulate this evaluation process by experienced radiologists.

In clinical setting, the degree of abnormality varies widely among different brain diseases. Ischemic infarction and tumor are diseases that often demonstrate large effect sizes, and MRI is considered one of the most effective diagnostic tools. At the other end of the spectrum are psychiatric diseases, for which MRI is not considered effective enough for routine clinical diagnosis. Dementia populations are located in the middle of the spectrum. Various dementia diseases with different causes and time courses are known to demonstrate brain atrophy in specific brain structures. However, this is compounded by the natural course of brain atrophy in aging brains and the coexistence of multiple pathology states unique to each patient, which would make the study

* Corresponding author at: Johns Hopkins University School of Medicine, Traylor 330, 720 Rutland Avenue, Baltimore, MD 21205, USA.
E-mail address: smori1@jhmi.edu (S. Mori).

http://dx.doi.org/10.1016/j.nicl.2016.09.008

of simple anatomy-pathology correlations challenging. Through past clinical experience and research, loose relationships between brain pathology and anatomical features have been established. For example, hippocampal atrophy is often used as an indicator of Alzheimer's disease, and frontotemporal dementia usually accompanies atrophy of the frontal and temporal lobes. However, such correlations may not be strong enough to be reliably perceived by subjective evaluations.

In conventional image analysis, the patient populations are first homogenized into specific dementia groups based on clinical symptoms (e.g., MCI, AD, etc.), and then, voxel-based analysis is performed to identify certain anatomical features that differentiate the population from a control group. The challenges faced by clinicians are fundamentally different as the patient population is inherently heterogeneous both in the anatomical and clinical domains, and the initial stratification of patients is often unknown. In addition, patients may have a heterogeneous "nature" and "degree" of pathologies that co-exist in multiple "locations." Thus, to support clinical image reading, it is an important effort to develop and test tools that can detect not only abnormalities (with respected to age-matched controls), but also apply past knowledge to individual patients and provide a quantitative and systematic evaluation of their anatomical states.

In the past, CBIR has been attempted for several radiological images, such as lung CT and mammography (Muller et al., 2004). For brain MRI, Coupe et al. (2012) devised a simultaneous segmentation and grading approach, which graded the test subjects based on several pre-selected structures, according to their similarity to the corresponding structures in the training subjects. This approach has been shown to provide respectable accuracy (up to 90% success rate) in discriminating AD patients and normal controls. The performance was superior or equivalent to other recent studies for AD classification, e.g., single-atlas based studies by Cuingnet et al. (2011) (88.5% accuracy) and Liu et al. (2012) (90.8% accuracy), or multi-atlas based studies by Koikkalainen et al. (2011) (86.0% accuracy) and Min et al. (2014)) (91.64% accuracy), just to name a few from the extensive literature about AD studies. The analysis in Coupe et al. (2012), however, was based on a hypothesis that the pre-selected structures (the hippocampus and the entorhinal cortex) carry key anatomical features for AD patients. This is a valid assumption for pre-stratified AD populations, but, for clinical patient populations with heterogeneous pathology, identifying key brain anatomical locations would be an important initial step, as the abnormality is region specific to each disease (or functional deficit) category.

This paper shares the same fundamental concept as the pioneering work by Coupe et al., but employs a unique framework of multiple-atlas brain segmentation to evaluate the anatomy of the entire brain. In atlas-based segmentation approaches (Collins et al., 1995; Dawant et al., 1999; Fischl et al., 2002; Joshi et al., 2004; Rohlfing et al., 2004), an atlas with pre-defined structures is warped to a patient image, thus transferring the structural definition for automated segmentation. In the multiple-atlas approach (Artaechevarria et al., 2009; Heckemann et al., 2006; Klein et al., 2005; Lotjonen et al., 2010; Warfield et al., 2004), there are multiple atlases that are all warped to a patient image. This leads to different segmentation results of a structure, followed by a fusion process (Langerak et al., 2010; Sabuncu et al., 2010; Warfield et al., 2004) to derive the best estimation of the structure. During the fusion, if all atlases receive equal weighting, majority voting prevails (Heckemann et al., 2006; Rohlfing et al., 2004). In more advanced approaches, each atlas receives a weighting based on anatomical similarity measures, such as the voxel intensity (Maes et al., 1997). By applying proper atlas-weighting, we expect to choose atlases with a similar anatomy and better registration accuracy, and thus, higher segmentation accuracy. Depending on algorithms, this operation is performed in a voxel-by-voxel or label-by-label manner. The content of the atlas library is the subject of various interesting questions. These include how many atlases are needed (Heckemann et al., 2006), whether they should be age-matched (Aljabar et al., 2009), or whether they should include pathological cases. In this study, we used the JHU multi-atlas library (Djamanakova et al., 2014; Wu et al., 2015) that contains images from healthy volunteers with a wide range of age, as well as images from patients, including mild cognitive impairment (MCI) and Alzheimer's disease (AD).

In typical multi-atlas segmentation studies, segmentation accuracy is usually the main interest of the studies, but in this paper, we focused on the atlas weighting as a measure of diagnostic voting from multiple atlases (called multi-atlas voting (MAV), hereafter); namely, if the AD-type of atlases are heavily weighted among all atlases for a structure of interest, the structure is judged as highly demented and the corresponding attributes (functional/diagnostic) from AD atlases are weighted heavily in the estimation of patient attributes. To test this idea, we first demonstrate the proof-of-principle using the age estimation, which can be validated by the known age. Then, we used MCI and AD cases from the Alzheimer's disease Neuroimaging Initiative (ADNI) to evaluate the performance in estimating both the functional deficits, such cognitive scores, and diagnostic categories (NC, MCI, or AD) of these patients. The accuracy was then measured by comparing the estimated and actual clinical data from the ADNI database. If the MAV estimation works, it would estimate AD patients as "highly demented" based on the anatomical features of medial temporal lobe structures , as the atrophy of these structures has been repeatedly identified as key discriminating features of AD in past morphometric studies. This paper describes the results of this unique MAV approach for direct estimation of the patients' diagnostic attributes in the context of Alzheimer's disease.

2. Methods and materials

2.1. Subjects

2.1.1. Subjects used for age estimation

The age-specific, multi-atlas dataset consisted of T1-weighted images of healthy controls from a pediatric population (4–12 yr, 20 atlases), a mid-age population (20–50 yr, 20 atlases), and an elderly population (60–80 yr, 20 atlases). Another 10 atlases from each age group were used as test subjects. The atlases are a subset of the MriCloud atlas repository (https://braingps.mricloud.org/atlasrepo), which were segmented to 289 structures with extensive manual correction. All images were acquired on Philips 3 T scanners, with image resolution in the range of $1.0 \times 1.0 \times 1.0$ mm to $1.0 \times 1.0 \times 1.2$ mm.

2.1.2. Subjects used for dementia estimation

The dementia-specific, multi-atlases consisted of T1-weighted images from the ADNI database (http://adni.loni.usc.edu/), with 20 atlases from the Alzheimer's disease (AD) population, 20 from the Mild Cognitive Impairment (MCI) population, and 20 from the normal elderly controls. Another 90 subjects ($n = 30$ in each group) chosen from ADNI as test subjects (Table 1). We estimated the diagnostic categories, as well as the functional attributes of the dementia patients, and we chose one of the most widely used cognitive scores—the Alzheimer's Disease Assessment Scale-cognitive subscales with 11 items (ADAS.11) (Llano et al., 2011)—as the functional measure. The patient information, including the ADAS.11 scores are summarized in Table 1. Other functional measurements, such as The Mini Mental State Examination (MMSE), can be calculated in the same way.

The ADNI data included data acquired from Philips, SIEMENS, and GE scanners at 1.5 T and 3 T. We used an even number of subjects from each protocol in each group (control, MCI, and AD) for both the training and testing datasets (Table 1). Our analysis, therefore, contains effects from image protocol differences. In our previous paper (Liang et al., 2015), we evaluated the protocol effects on our pipeline, and found that the pipeline could robustly detect age-dependent anatomical changes regardless of the wide variety of protocols, including the different scanner fields and manufacturers. We evaluated the effect of scan protocol on

Table 1
ADNI data used for diagnosis estimation. Abbreviations: P1.5 – Philips 1.5 T; P3 – Philips 3 T; S1.5 – Siemens 1.5 T; S3 – Siemens 3 T; G1.5 – GE 1.5 T; G3 – GE 3 T.

Group	No.	Usage	Age (years)	Diagnosis (ADAS.11)	Number of subjects from P1.5/P3/S1.5/S3/G1.5/G3
Control	20	Atlas	70.8 ± 8.3	4.53 ± 2.20	3/4/3/4/3/3
Control	30	Test	71.6 ± 2.5	6.57 ± 3.49	5/5/5/5/5/5
MCI	20	Atlas	73.1 ± 9.5	11.75 ± 2.81	3/4/3/4/3/3
MCI	30	Test	71.4 ± 8.7	12.78 ± 4.07	5/5/5/5/5/5
AD	20	Atlas	70.7 ± 11.0	17.05 ± 3.99	3/4/3/4/3/3
AD	30	Test	69.7 ± 12.3	20.67 ± 5.05	5/5/5/5/5/5

the estimation of patient attributes, as described in Section 2.3.4. We believe the inclusion of various MPRAGE protocols (all provided by the manufacturers) in this study was highly important to ensure that the observed biological effects would not be erased in practice when different imaging protocols are used.

2.2. Algorithms

2.2.1. Multi-atlas segmentation framework

In multi-atlas based segmentation, the parcellation profiles of the target image from each atlas, after registration, are combined according to certain atlas-weighting and fusion schemes. The registration in this study was achieved first by affine transformation, and then iterative Large Deformation Diffeomorphic Metric Mapping (LDDMM) (Christensen et al., 1996; Grenander and Miller, 1998; Miller et al., 1993), along with iterative inhomogeneity corrections. The concept of multi-atlas approach assumes that the warping of a single atlas to the target image is not always perfect. Although LDDMM ensures preserved topology and invertible transformation from an atlas to the target, if the anatomical differences are large, it could lead to a highly stretched transformation.

Let I_T be the target image, I_A^i ($i=1,2,\ldots,N$) be the atlas images after warping to the target image, and L_A^i be the label images associated with the warped atlases. We used a weighted voting approach (Artaechevarria et al., 2009; Isgum et al., 2009; Sabuncu et al., 2010) for label fusion:

$$\hat{p}(l|x,I_T)=\sum_{i=1}^{N} w_A^i(x)\cdot p\left(l|x,I_A^i\right) \tag{1}$$

where $\hat{p}(l|x,I_T)$ is the estimated probability of voxel x being labeled l in the target image, and $l=1,2,\ldots,L$, with L the total number of labels. $p(l|x,I_A^i)$ is the probability of voxel x being labeled as l in the warped atlas, with $p(l|x,I_A^i)=1$ when $L_A^i(x)=l$ and $p(l|x,I_A^i)=0$ otherwise. $w_A^i(x)$ represents the atlas-weighting term that measures the similarity between the target and atlas i at voxel x, with $\sum_{i=1}^{N} w_A^i(x)=1$. When $p(l|x,I_A^i)$ is binary, as it is used in this study, Eq. (1) reduces to a weighted majority voting scheme. The atlas-weighting used in this study is described in the next section. The final segmentation can be obtained by the Bayes maximum a posteriori (MAP) estimation, $L_T(x)=\mathrm{argmax}_{l\in\{1,\ldots,L\}}\,\hat{p}(l|x,I_T)$.

2.2.2. Atlas-weighting strategy

Atlas-weighting is the key component in MAV-based disease estimation. We assign atlas-weightings to each individual structure, based on the intensity similarity, on a label-by-label basis, as opposed to a voxel-by-voxel based approach. The similarity is measured based on the local intensity match along the boundary of each structural label between the target and the warped atlases. Namely, a tentative, "atlas-specific" boundary is casted from the atlas to the target first and the degree of matching, based on the similarity of the voxel intensity along the atlas-specific boundary, is evaluated. This process is repeated for each atlas and the weighting was judged based on similarity. We choose the boundary voxels rather than all voxels in the label, assuming the image intensities inside the structure are relatively homogeneous and the boundary voxels are more sensitive to the structural similarity.

Let $N_x=[x_1,x_2,\ldots,x_K]$ be a vector of voxels in a local neighborhood patch of radius $r\times r\times r$ centered on a boundary voxel x, then the similarity measure $s_A^i(x)$ of a warped atlas i is computed by

$$s_A^i(x)=corr\left(I_A^i(N_x),I_T(N_x)\right) \tag{2}$$

where $corr(\cdot)$ is the Pearson correlation coefficient $corr(I_A^i(N_x),I_T(N_x))=\frac{E[(I_A^i(N_x)-\mu(I_A^i(N_x)))(I_T(N_x)-\mu(I_T(N_x)))]}{\sigma(I_A^i(N_x))\sigma(I_T(N_x))}$ with E, μ, and σ being the expectation, mean, and standard deviation operations, respectively. $s_A^i(x)$ is a signed quantity with negative values indicating anti-correlation. We assume anti-correlations rarely occur in practice, as we use the atlas and target images from the same modality, which are sufficiently co-registered to each other. In case of anti-correlation, a negative $s_A^i(x)$ value would indicate very low similarity.

Since the similarity is measured based on the warped atlases that are already transformed through a large deformation to match the target image, in order to properly weigh the patient attributes that are associated with the un-deformed atlases in their native space, we include a deformation cost in the atlas-weighting. The LDDMM provides a structure-preserving diffeomorphism φ between the atlas and target images by solving $\dot{\varphi}_t=v_t\circ\varphi_t$, $t\in[0,1]$, with $v_0=\dot{\varphi}_0$, and minimizing $\|I_T-I_A\circ\varphi_1^{-1}\|^2+\int_0^1\|v_t\|_V^2dt$, where v_t is the time-dependent velocity vector field of the flow of the deformation, φ_t is the diffeomorphism at time t, $\dot{\varphi}_t$ denotes the first-order differentiation of φ_t, and $\int_0^1\|v_t\|_V^2dt$ represents the integration of the norm of v_t over the entire velocity field, V (Christensen et al., 1996; Grenander and Miller, 1998; Miller et al., 1993, 2015). The deformation energy at individual voxels can be approximated by the determinant of the Jacobian matrix of the LDDMM transformation $J(x)$. We calculate the deformation cost using a negative exponential function of this quantity, $\alpha_A^i(x)=\exp(-J_A^i(x))$, for each atlas I_A^i at voxel x, such that the larger the deformation, the smaller the α_A^i. In the boundary-based atlas-weighting, we obtain $\alpha_A^i(x)$ at each boundary voxel along an atlas label, and the dot product of $s_A^i(x)\cdot\alpha_A^i(x)$ summed over all boundary voxels gives the weighting of atlas I_A^i for this label.

$$w_A^i(l)=\sum_{x\in b_A^i(l)} s_A^i(x)\cdot\alpha_A^i(x) \tag{3}$$

where $w_A^i(l)$ is the atlas-weighting of label l in atlas i, and $b_A^i(l)$ denotes the boundary of label l in the warped atlas i.

There are slight differences in the use of atlas-weighting for segmentation and MAV purposes. In segmentation, the atlas fusion is performed per spatial location and all voxels within the same label of an atlas receive the same weighting; whereas, for the MAV-based direct estimation of patient attributes, the voting is performed per structure. In both cases, the atlas-weightings are normalized such that $\sum_{i=1}^{N} w_A^i(l)=1$. In addition, we used a combined weighting based on the similarity of the deformed atlas and the deformation cost for direct estimation of patient attributes; whereas, for segmentation purposes, we used only the similarity of the deformed atlas in atlas-weighting, $w_A^i(l)=\sum_{x\in b_A^i(l)} s_A^i(x)$, as the deformation cost is not related to the segmentation.

2.2.3. MAV-based patient attribute estimation

Given the demographic or clinical information, $D(I_A^i)$, that is associated with atlas i, the same type of information about the target subject can be inferred by

$$D(I_T|l) = \sum_{i=1}^{N} D\left(I_A^i\right) \cdot w_A^i(l) \quad (4)$$

where $D(I_T|l)$ is the demographic or clinical estimation of structure l in a target subject. In the age estimation test, $D(\cdot)$ will be the age measure, and in the functional estimation of the dementia patients, $D(\cdot)$ will be the ADAS.11 score.

We can also estimate the diagnostic categories of the dementia patients based on the probability of the target subject belonging to normal elderly, MCI, or AD groups. The can be achieved by summing over the weightings of the atlases associated with the specific diagnostic groups, after normalizing the sum of all atlas weights to unity.

$$p(G_j|I_T, l) = \sum_{i \in G_j} w_A^i(l) \quad (5)$$

where $p(G_j|I_T, l)$ is the probability of the target belonging to atlas group G_j in terms of label l, with $j = 1, 2, \ldots, J$ (the number of diagnostic groups).

2.3. Performance evaluation and statistical tests

2.3.1. Correlation between the estimated and clinically measured patient attributes

In the age estimation test, the ages estimated from the test subjects (n = 30) on individual structures were compared with the subjects' actual ages by linear regression. The estimation of functional deficits in the dementia population was similarly evaluated by linear regression between the estimated ADAS.11 scores and the clinically measured ADAS.11 in the ADNI subjects (n = 90). The R^2 was used to evaluate the goodness-of-fit of the linear regression, and the p-value from the t-statistics was used to evaluate the significance of linear regression with False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) correction. The linear correlations between volumes of each structure and age or ADAS.11 were also computed for comparison. The volumes were obtained using a multi-atlas segmentation pipeline, as described in Section 2.2.1 and the proposed atlas-weighting in Section 2.2.2.

2.3.2. Group differences

To assess the significance of the estimated ADAS.11 among ADNI groups, a one-way analysis of variance (ANOVA) was performed among the AD/MCI/NC test subjects (n = 30 each), and the p-values from ANOVA tests were obtained and corrected by FDR for multiple ROI comparisons. The same ANOVA test was also performed on the volumetric measurements.

2.3.3. Feature extraction and classification accuracy

In addition to the regional presentation of patient attributes estimated from the MAV diagram, the regional features can also be combined to classify the disease categories. We compared the classification performance using the volumetric measurements, estimated ADAS.11 scores, and the estimated diagnostic category probabilities based on Eq. (5). In order to use the three types of categorical probabilities (AD/MC/NC) for feature selection and classification purposes, we integrated them into a single dementia probability measurement with a simple linear combination: $p(Dementia|I_T, l) = p(G_{AD}|I_T, l) \times 1.0 + p(G_{MCI}|I_T, l) \times 0.5 + p(G_{NC}|I_T, l) \times 0.0$, for structure l of target image I_T.

To extract the most discriminative features from the high-dimensional feature vector (volumes or dementia probabilities estimated from 289 brain structures), we tested two approaches using the training data (atlases, n = 20 per group): 1) the top one or top 20 structures that showed the most prominent group difference, based on the p-values from the ANOVA tests; 2) the LASSO method (least absolute shrinkage and selection operator) (Tibshirani, 1996), which is a regression analysis method that selects the best subset of variables to enhance prediction accuracy. In our study, we performed LASSO using a regression model between the pre-determined diagnostic category (response variable) and the volumes, estimated ADAS.11 scores, or dementia probabilities from each structure (covariates). The LASSO method determines the optimal number of structures when the mean square fitting error is minimum. The regional features selected by the top one or top 20 criteria or LASSO were then fed into a linear discriminant analysis (LDA) classifier, using a leave-one-out cross validation approach on the ADNI test subjects (n = 90). The sensitivity, specificity, and overall accuracy of two-category (AD/NC) or three-category (AD/MCI/NC) classifications were evaluated. The LASSO and LDA were performed using R packages (https://cran.r-project.org/).

2.3.4. Evaluation of protocol effect

We evaluated the effect of scan protocol by including the protocol type (six types of protocols used in ADNI data acquisition) as another factor in addition to the estimated ADAS.11 scores, and tested its significance among the three groups, using two-way ANOVA followed by FDR correction. The protocol effect was statistically significant only in two of the 289 brain segments (left fusiform gyrus and left subcortical white matter of the inferior temporal gyrus).

3. Results

3.1. Framework of multi-atlas voting

The concepts of the two approaches are summarized in Fig. 1. The first approach that is illustrated in the blue dashed box is based on MAV, as described above. In this approach, the patient attributes (age as an example) are obtained directly from the process of the multi-atlas pipeline and the resultant segmentation is merely a proof of procedural accuracy (as long as the segmentation is accurately performed, the segmentation results are discarded). The second approach, as illustrated in the yellow dashed box, is a more conventional method, in which the segmentation results (e.g., volumes) are compared with population-based regression for indirect estimation of the patient's age. The population-based regression between the volume and patient attributes must be established beforehand as a priori knowledge for the estimation of any new patients.

3.2. Testing of regional feature estimation using age

We used age estimation as a proof-of-concept of the MAV approach. Because we knew the exact age of each subject, we could evaluate the accuracy of the age estimation. The estimation was performed in each test subject (n = 30) on a structure-by-structure basis according to Eq. (4). The linear regression between the estimated ages and actual ages showed significant linear correlation (family-wise $p < 0.01$) in a majority of the structures (214 out of 289). Fig. 2 shows the correlation plots in several cortical, subcortical gray matter, and white matter regions. The subcortical structures and deep white matter structures demonstrated high correlation between the estimated age (y-axes) and the actual age (x-axes), with R^2 values around 0.7. The correlation with cortical structures was relatively weak, with R^2 around 0.3–0.5. The R^2 values and the slopes of linear regression were mapped to the T1-weighted images, and masked by a family-wise p-value threshold of 0.01 (Fig. 3). The R^2 maps indicated that the age estimation is most precise in the subcortical gray matter, the anterior deep white matter, and the cerebellum. Some peripheral white matter tracts and gyri in the posterior and superior brain did not show significant correlation. The slopes of the linear regression, which represent the systematic bias between

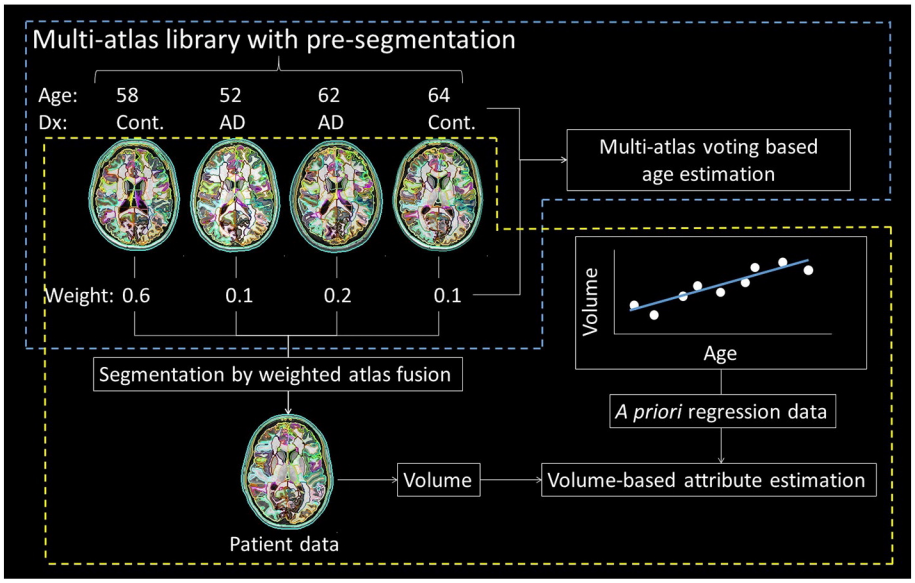

Fig. 1. A schematic showing the concepts of multi-atlas voting (MAV)-based analysis and conventional region-of-interest (ROI)-based analysis. In the MAV approach (dashed blue box), the similarity between the patient images and the atlases is measured based on the image features, which is then used to weigh the attributes (age as an example) associated with the multiple atlases to obtain a weighted estimation of the patient's attribute. In comparison, in ROI-based analysis (dashed yellow box), the multi-atlases are used to segment the image, and the volumes or intensities of the ROIs are used to estimate the patient's attribute in an indirect manner, which relies on a priori regression data between the volume and patient attributes (age as an example here).

the estimated and actual ages, suggested high estimation accuracy in the thalamus and midbrain structures and a higher degree of bias in the peripheral structures. Fig. 3C demonstrates the regional age estimation across the brain in several representative individuals at 7, 22, 41, 63, and 80 years of age. A clear increase in the estimated age was observed from the pediatric to the elderly brains. Regional variations were present, especially in the middle-age range.

We compared the age estimation by the MAV with a simple volume-based approach. The linear correlation between volumes and ages reached significance (family-wise $p < 0.01$) in 174 out of 289 structures.

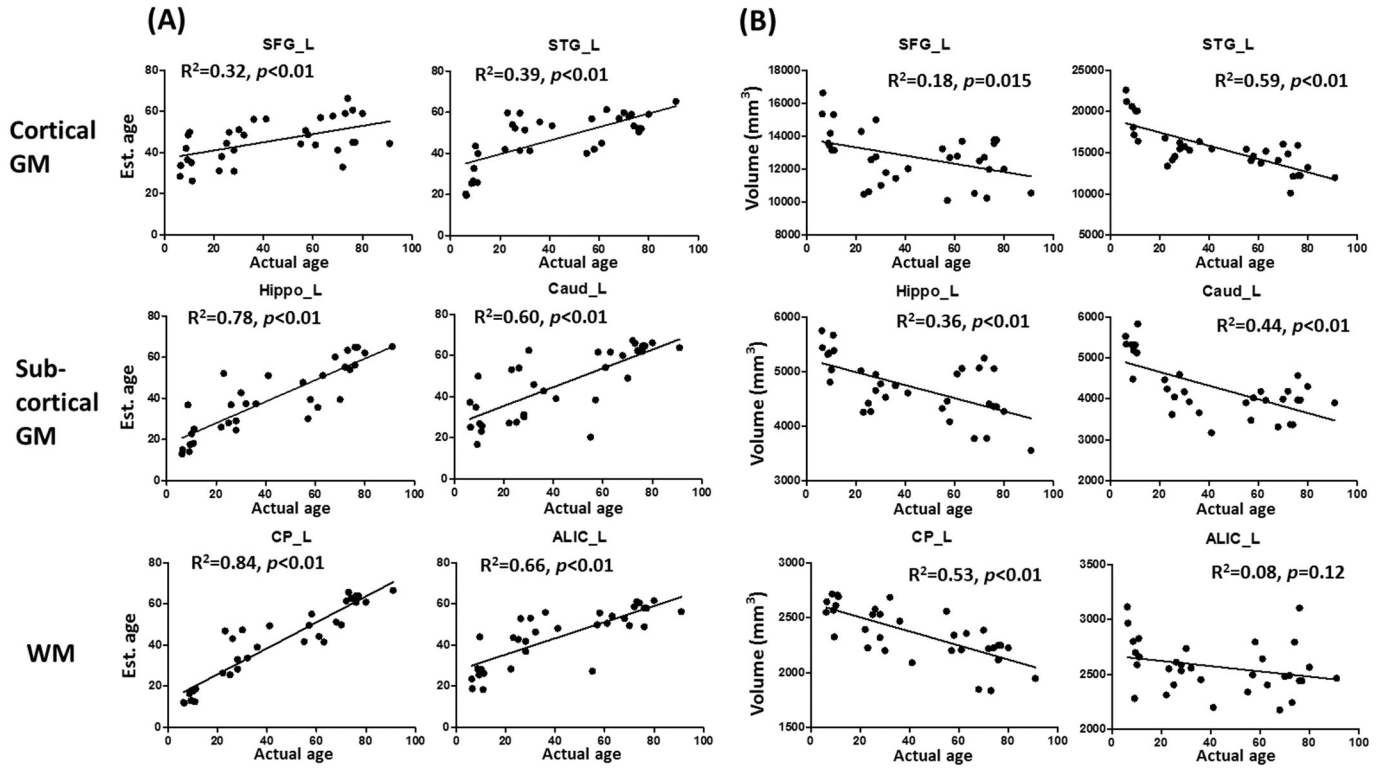

Fig. 2. (A) Linear regression between the estimated ages (y-axes) and actual ages (x-axes) of 30 test subjects in several cortical, subcortical gray matter, and white matter structures. (B) Linear regression between the structural volumes (y-axes) and ages (x-axes) in the same structures as in (A). The R^2 and p-values of the linear regression are denoted in each graph. Abbreviations: SFG_L- left superior frontal gyrus; STG_L- left superior temporal gyrus; Hippo_L- left hippocampus; Caud_L- left caudate; CP_L: left cerebral peduncle; ALIC_L- left anterior limb of the internal capsule.

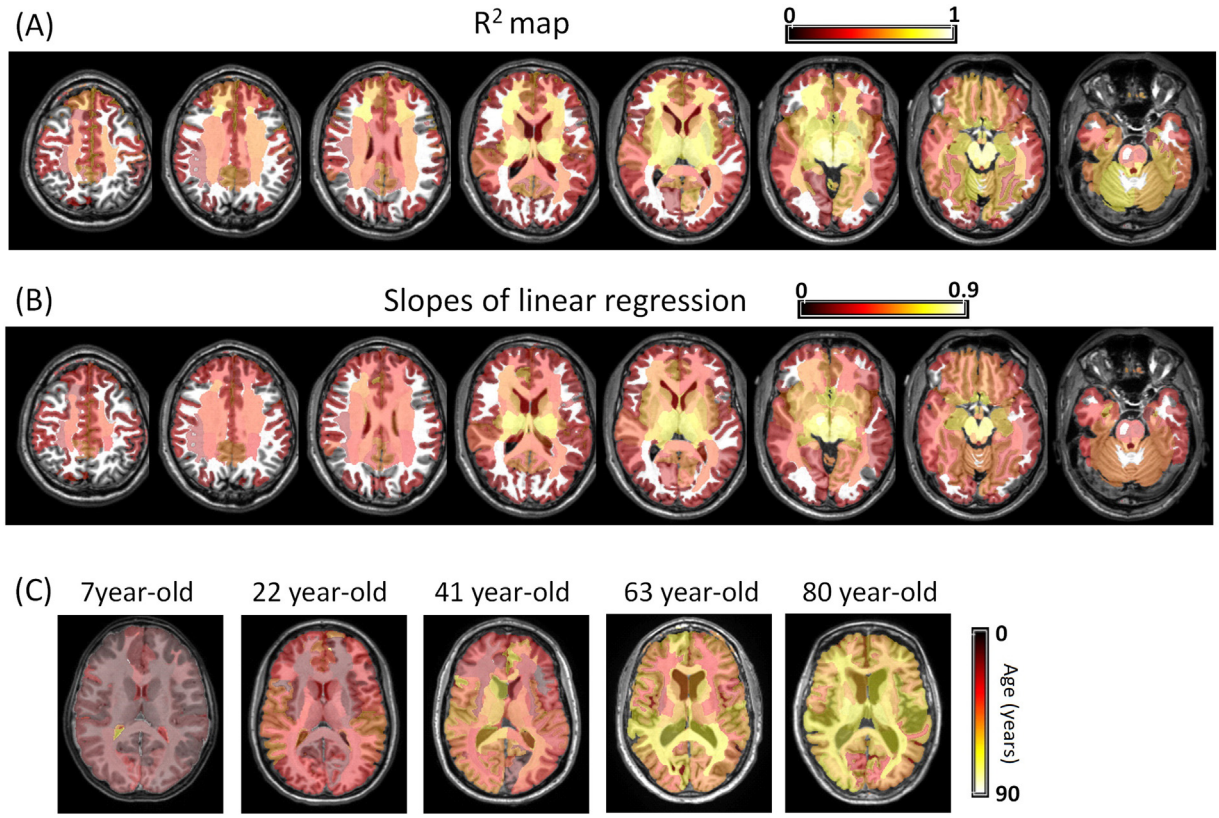

Fig. 3. Whole-brain mapping of the R^2 and linear correlation coefficients of the linear regression between the estimated age and actual age in each structure, overlaid on a T1-weighted image. Only structures with significant linear regression (family-wise p-value < 0.01) are shown. Dark red indicates low R^2 or correlation coefficients, and the bright color indicates high values.

Fig. 2B compares the volume-to-age linear correlation in the same structures as the MAV-based correlation plots in Fig. 2A. The R^2 values of volume-based and multi-atlas-based linear regression are directly compared in all 289 structures in Fig. 4. In the subcortical gray matter and deep white matter, the MAV-based age estimation outperformed volume-based estimation; whereas, in the cortical structures, the R^2 of volume-based correlation was relatively higher. With the MAV-based approach, 48 structures reached $R^2 > 0.7$, among which several subcortical gray matter, deep white matter, and ventricle structures showed R^2 values of 0.8 or higher, while there were only six structures that reached $R^2 > 0.7$ with the volume-based approach. In 200 of the 289 structures, the MAV-based age prediction gave higher R^2 values than the volume approach.

3.3. Estimation of the functional states in the dementia population

We estimated the cognitive assessment (ADAS.11 score) of the ADNI subjects using the disease-specific, multi-atlas library according to Eq.

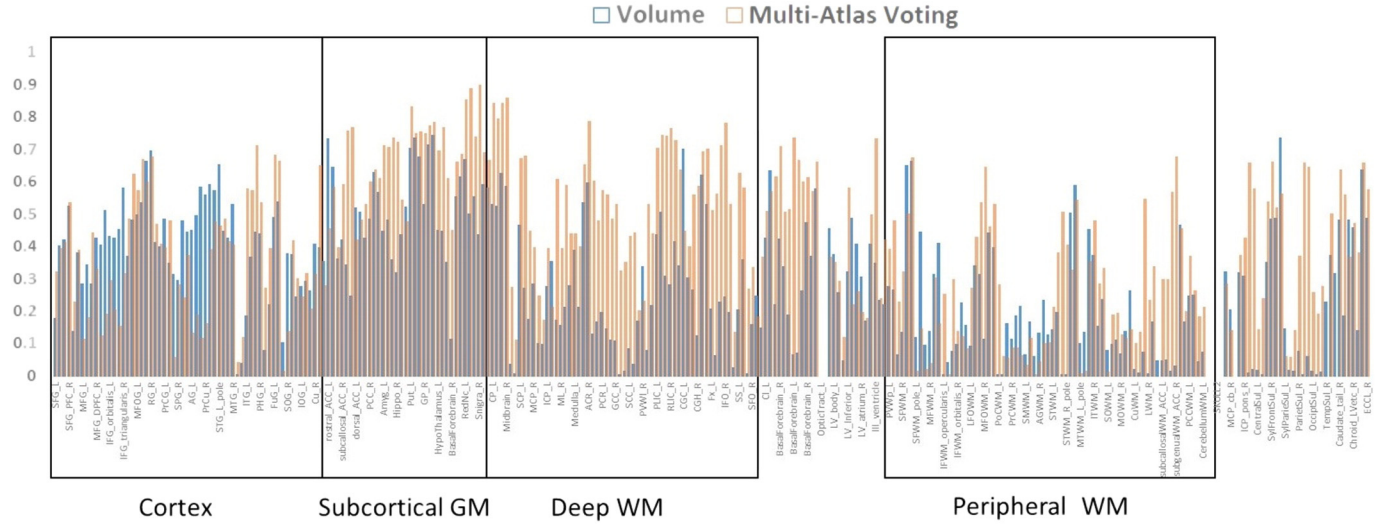

Fig. 4. R^2 of the linear regression between the structural volume and age (blue bars), compared to the R^2 of the linear regression between the MAV-based estimation and age (red bars), in 289 structures over the whole brain.

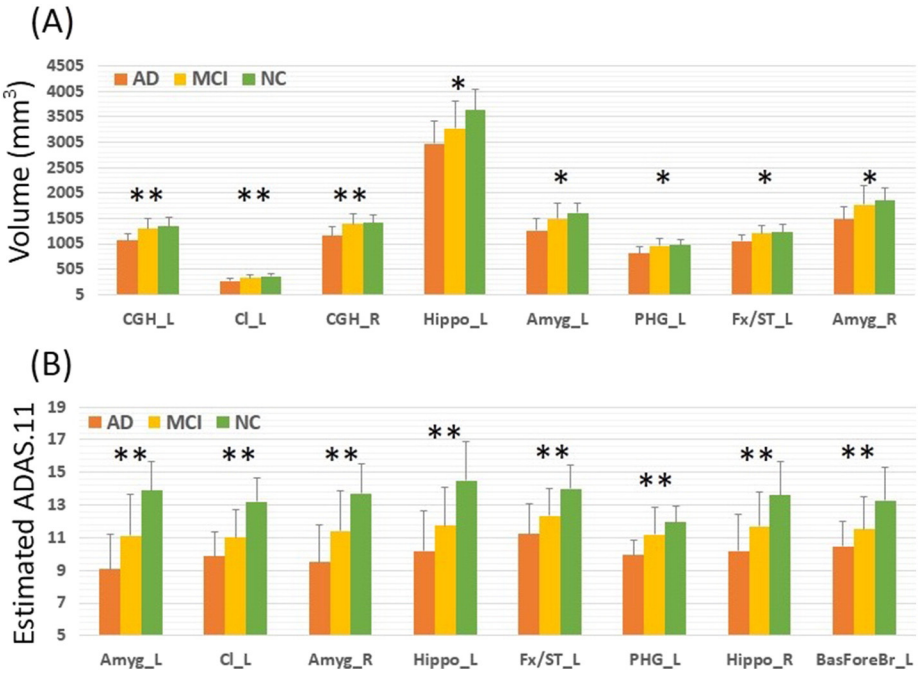

Fig. 5. Volumes (A) and MAV-based estimation of ADAS.11 scores (B) from the AD, MCI, and control test subjects (n = 30 in each group, presented as group mean ± standard deviation), in the structures that showed the most significant group difference. The order of the structures was determined based on their p-values (from low to high) from by one-way ANOVA followed by FDR correction. *$p < 0.001$, **$p < 1 \times 10^{-5}$. Abbreviations: CGH - cingulum (hippocampal part); CL – claustrum; Hippo – hippocampus, Amyg – amygdala; PHG - parahippocampal gyrus; Fx/ST – fornix/stria terminalis; BasForeBr – basal forebrain.

(4). We first examined the group average ADAS.11 scores estimated in the normal elderly, MCI, and AD test subjects (n = 30 each). Fig. 5A–B shows the top eight brain regions with the largest group differences in terms of the volumes (Fig. 5A) or estimated ADAS.11 (Fig. 5B), ranked according to the p-values from one-way ANOVA (after FDR correction). The two charts shared similarities: six of the eight detected structures

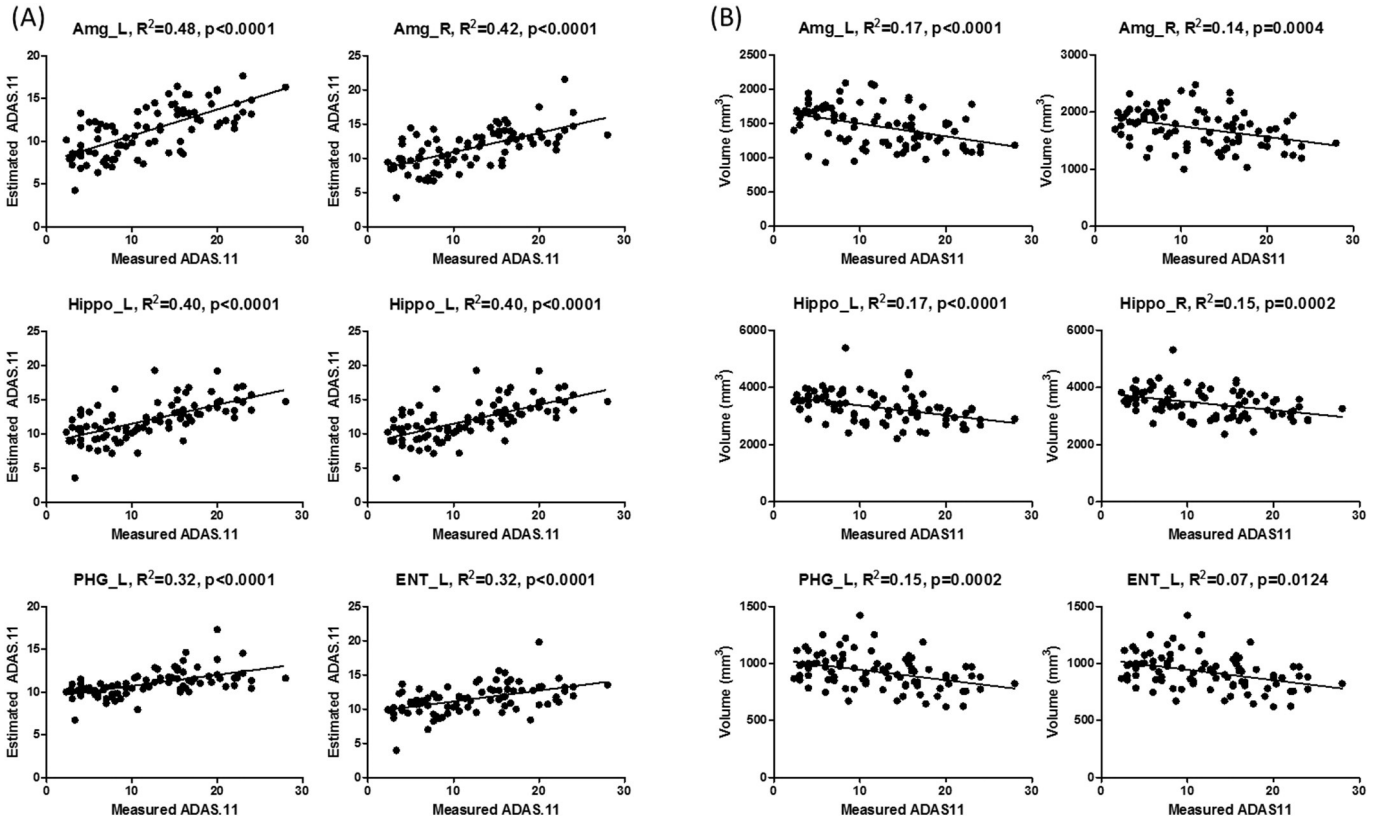

Fig. 6. (A) Linear regression between the estimated ADAS.11 scores (y-axes) and clinically measured scores (x-axes) of 90 test subjects in several structures with the highest R^2, including the left and right hippocampus, amygdala, left parahippocampal gyrus and left entorhinal cortex. (B) Linear regression between the structural volumes (y-axes) and ADAS.11 score (x-axes) in the same structures. The R^2 and p-values of the linear regression are denoted in each graph.

were limbic gray matter structures, including the left and right amygdala, the left hippocampus, and the left parahippocampal gyrus, as well as limbic white matter structures, including the cingulum and the fornix/stria terminalis. Overall, fewer structures showed significant group differences (family-wise $p < 0.01$) using volumetric measurement (18 out of 289 structures), compared to those using estimated ADAS.11 (67 out of 289 structures).

Linear correlation between the estimated ADAS.11 scores (y-axes) and the clinically measured scores (x-axes) was plotted in Fig. 6A for several structures that showed the strong correlations according to the R^2. The left and right amygdala showed highest correlation ($R^2 = 0.48$ and 0.42), followed by the left and right hippocampus ($R^2 = 0.4$). These structures are known to be most affected in Alzheimer's patients. In comparison, the correlations between volumes and measured ADAS.11 were lower in these structures ($R^2 < 0.2$, Fig. 6B); whereas the highest volume-based R^2 were found in the left claustrum ($R^2 = 0.29$). The R^2 maps in Fig. 7A showed significant correlations between the estimated and measured ADAS.11 (familywise $p < 0.01$) in multiple subcortical gray matter structures, including the caudate, the thalamus, and the basal forebrain; the white matter structures, including the internal capsule, the cingulum, and the corpus callosum; the lateral ventricle; and selected cortical regions, especially in the temporal lobes. Lateralization was observed in some structures toward the left hemisphere. The slopes of the linear regression (Fig. 7B) were highest in the hippocampus and the inferior lateral ventricle.

3.4. Testing of disease classification

The diagnostic categories (AD/MCI/NC) were predicted using three types of MRI markers – the conventional structural volumes, the dementia probabilities (as described in Section 2.3.3) and the ADAS.11 scores estimated from MAV. The results are summarized in Table 2. In this analysis, the classification accuracy was tested using the most discriminating structures, based on their significance in differentiating the AD/MCI/NC groups, or based on the LASSO analysis using the training (atlas) data. For the group difference ranking, we used the top one or top 20 structures of the total of 289 structures from both the volume or dementia probability markers for classification, while LASSO determined an optimal number of 22 structures from the volume marker, 40 structures from dementia probability estimations, and 26 structures from the estimated ADAS.11 scores. Table 3 lists the top 20 structures selected from each approach. As expected, the limbic structures in the medial temporal lobes dominated the list. For the volumetric markers, the left cingulum, which gave the most significant group differences, was used as the single feature for classifying AD/MCI/NC, which resulted in an overall accuracy of 0.53. The combination of the features based on the ranking or LASSO improved the overall accuracy to 0.58 and 0.61, respectively. The use of dementia probability estimated from MAV further improved the overall accuracy to 0.63, 0.69, and 0.82 for a single structure (left amygdala), combinations by ranking, and LASSO, respectively. The classification performance using the estimated ADAS.11 scores were in-between the performances of the volume and dementia probability markers, with an accuracy of 0.56, 0.61, and 0.69 for a single structure (left amygdala), the group difference ranking, and LASSO, respectively. In addition, we performed classification using the clinically measured ADAS.11 scores, which resulted in an accuracy of 0.79. If the estimation was limited to AD vs NC (MCI data excluded), the accuracy was 0.85–0.88 with volumetric markers, 0.9–0.92 with dementia probability estimation, and 0.88–0.92 with the estimated ADAS.11 scores. The AD/NC classification was 100% correct using the clinically measured ADAS.11 scores.

4. Discussion

4.1. Concept of multi-atlas voting for disease estimation

MRI atlases are commonly used for automated image segmentation (Chupin et al., 2009; Collins et al., 1995; Dawant et al., 1999; Fischl et al., 2002; Joshi et al., 2004; Rohlfing et al., 2004), and provide pre-segmented maps as a priori knowledge about the shapes and locations of the structures to guide the segmentation. The use of multiple atlases yields robust and accurate segmentation (Artaechevarria et al., 2009; Heckemann et al., 2006; Jia et al., 2012; Klein et al., 2005; Lotjonen et al., 2010), as the rich anatomical information from multiple atlases offers the flexibility to accommodate the diverse anatomy of the patient population. The end-goal of the atlas- or multi-atlas-based approaches is typically to obtain accurate segmentation, from which information about volumes, intensities, or shapes of the segmented structures can be extracted (Heckemann et al., 2008; Kloppel et al., 2008; Mori et al., 2013; Tustison et al., 2014). Much of the previous effort has been focused on improving the segmentation accuracy through advanced image registration (Gholipour et al., 2007; Klein et al., 2009; Lotjonen et al., 2009) or atlas weighting and fusion strategies (Langerak et al., 2010; Sabuncu et al., 2010; Tang et al., 2013; van Rikxoort et al., 2010; Wang et al., 2013; Warfield et al., 2004).

An interesting aspect of these studies is that the determination of the structural volumes is usually NOT the ultimate goal of the study; instead, the raw volume numbers must be interpreted to extract clinically useful information. For example, the structural volumes are compared between control and patient groups (and thus, can serve as a biomarker for diagnosis) or correlate with brain function measures (and thus, can be used to infer the functional loss). Therefore, the volume information

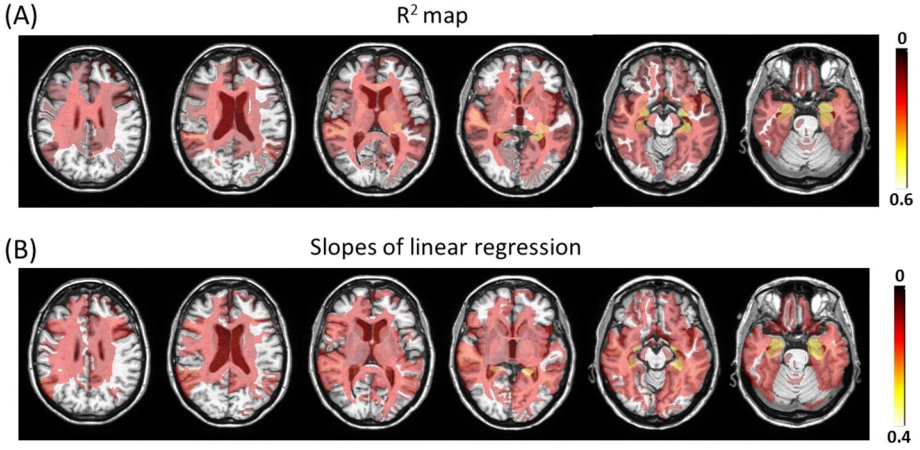

Fig. 7. Whole-brain mapping of the R^2 (A) and linear correlation coefficients (B) of the linear regression between the estimated ADAS.11 and clinically measured score in each structure, overlaid on a T1-weighted image. Only structures with significant linear regression (family-wise p-value < 0.01) are shown.

Table 2
The sensitivity, specificity, and overall accuracy of the LDA classification results. Three types of biomarkers were evaluated — the structural volumes, the dementia probabilities, and the ADAS.11 score estimated based on the multi-atlas voting approach. Three feature extraction approaches based on the group difference rank (top one and top 20) and LASSO were tested, for each of the markers. We performed both three-group (AD/MCI/NC) and two-group (AD/NC) classification. In addition, we tested the classification results using the clinically measured ADAS.11 score. In the three-group classification, sensitivity and specificity were determined for each group. The overall accuracy is the percentage of true-positive plus true-negatives.

Measurement	Feature selection	Classification groups	Sensitivity	Specificity	Overall accuracy
Volume	Single structure (left cingulum)	AD/MCI/NC	0.73/0.27/0.60 (AD/MCI/NC)	0.80/0.73/0.77	0.53
		AD/NC	0.80	0.90	0.85
	Group difference rank (top 20)	AD/MCI/NC	0.77/0.47/0.50	0.80/0.75/0.82	0.58
		AD/NC	0.83	0.83	0.83
	LASSO (22 structures)	AD/MCI/NC	0.70/0.43/0.70	0.87/0.73/0.82	0.61
		AD/NC	0.93	0.83	0.88
Dementia probability	Single structure (left amygdala)	AD/MCI/NC	0.83/0.40/0.67	0.80/0.80/0.85	0.63
		AD/NC	0.97	0.83	0.90
	Group difference rank (top 20)	AD/MCI/NC	0.63/0.63/0.80	0.87/0.77/0.90	0.69
		AD/NC	0.93	0.90	0.92
	LASSO (40 structures)	AD/MCI/NC	0.83/0.73/0.90	0.95/0.87/0.92	0.82
		AD/NC	0.87	0.93	0.90
Estimated ADAS.11	Single structure (left Amygdala)	AD/MCI/NC	0.73/0.27/0.67	0.83/0.72/0.78	0.56
		AD/NC	0.97	0.83	0.90
	Group difference rank (top 20)	AD/MCI/NC	0.67/0.47/0.70	0.83/0.73/0.85	0.61
		AD/NC	0.90	0.93	0.92
	LASSO	AD/MCI/NC	0.77/0.63/0.67	0.92/0.73/0.88	0.69
		AD/NC	0.93	0.87	0.90
Measured ADAS.11	None	AD/MCI/NC	0.73/0.63/1.0	0.88/0.87/0.93	0.79
		AD/NC	1.00	1.00	1.00

is an intermediate marker to infer more clinically meaningful information about the patients, such as diagnosis, prognosis, and functional risk factors.

The proposed MAV-based approach tries to directly estimate the clinically meaningful information from the knowledge database (atlas libraries) without going through the volume measurement step, which is naturally incorporated in the multi-atlas selection processes. The approach assumes that certain individual traits, such as demographics, functions, and diagnostics, have specific anatomical signatures (shape and intensity) at certain anatomical locations. Through the searching (weighting) for proper atlases based on anatomical similarity, the MAV approach attempts to retrieve non-image individual traits. For the anatomical similarity criteria, intensity-based atlas-weighting is widely used, e.g., the intensity differences, cross-correlation, or mutual information (Maes et al., 1997). Shaped-based averaging (Rohlfing and Maurer, 2007) is also an option, which requires an initial segmentation of the target image. The deformation energy of transformation between the atlases and targets can also be used (Rohlfing et al., 2004), as

Table 3
The structures selected for classification from the volumetric marker and dementia probability estimation, based on the group difference rank or LASSO method. Note that, in the group difference rank, the top 20 structures were chosen for classification; while LASSO method used 22 structures from the volumetric measurements and 40 structures from the dementia probabilities, and only the first 20 structures with the highest regression coefficients (absolute value) are shown here.

	Volume-based		Dementia probability-based	
	Group difference rank	LASSO	Group difference rank	LASSO
1	Left cingulum (hippocampal part)	Left claustrum	Left amygdala	Left parahippocampal gyrus
2	Left claustrum	Right claustrum	Left parahippocampal gyrus	Right basal forebrain
3	Right cingulum (hippocampal part)	Left pontine crossing tract	Left hippocampus	Right hippocampus
4	Left hippocampus	Left cingulum (hippocampal part)	Right amygdala	Left claustrum
5	Left amygdala	Right caudate tail	Right hippocampus	Left genu of corpus callosum
6	Left parahippocampal gyrus	Left hippocampus	Left claustrum	Right entorhinal cortex
7	Left fornix/stria terminalis	Right sylvian fissure and posterior insular sulcus	Left basal forebrain	Right subcortical white matter of the rostral anterior cingulate cortex
8	Right amygdala	Left parahippocampal gyrus	Left fornix/stria terminalis	Left amygdala
9	Right claustrum	Left gyrus rectus	Right fimbria	Right subcortical white matter of the fusiform gyrus
10	Right hippocampus	Left fornix/stria terminalis	Left ECCL	Left hippocampus
11	Right parahippocampal gyrus	Right cingulum (hippocampal part)	Left entorhinal cortex	Left sylvian fissure into supramarginal gyrus
12	Right fornix/stria terminalis	Left subcortical white matter of the lingual gyrus	Right basal forebrain	Left pontine crossing tract
13	Left cerebral peduncle	Left anterior part of the periventricular white matter	Right caudate tail	Right fimbria
14	Right inferior fronto-occipital fasciculus	Left occipital lateral ventricle	Right entorhinal cortex	Left fimbria
15	Right pontine crossing tract	Left superior frontal gyrus/pole	Right claustrum	Left basal forebrain
16	Left basal forebrain	Right inferior fronto-occipital fasciculus	Left cingulum (hippocampal part)	Left retrolenticular part of internal capsule
17	Left pontine crossing tract	Left parietal sulci	Left caudate tail	Left entorhinal cortex
18	Left inferior cerebellar peduncle/pons	Right hippocampus	Left genu of corpus callosum	Right caudate
19	Right substantia nigra	Left subcortical white matter of the middle frontal gyru	Right fornix/stria terminalis	Right pontine crossing tract
20	Right thalamus	Right inferior lateral ventricle	Right inferior fronto-occipital fasciculus	Right lateral part of the periventricular white matte

less deformation indicates higher similarity between the images in the native space. The similarity-based atlas-weighting can be evaluated on a global scale, such as the whole brain (Aljabar et al., 2009), or on localized scales, such as the voxels and structures (Wu et al., 2007). Artaechevarria et al. showed that locally defined weights improved the segmentation compared to the global approach (Artaechevarria et al., 2009).

In this study, we computed the atlas-weighting on a structure-by-structure basis to reflect the local pathology. Our strategy is to focus on the boundary voxels of each structural label, trying to search for images with similar anatomical features for not only the shape of the structure of interest, but also for the surrounding anatomical features. For example, the medial, lateral, and dorsal surfaces of the hippocampus are surrounded by the ventricles. In normal healthy brains, large portions of the ventricles in the dorsal and lateral surfaces are closed (invisible on MRI with 1 mm resolution) and the adjacent white matter tissues seem attached to the hippocampus, while these ventricle spaces enlarge and become visible in patients with severe brain atrophy. The boundary-based similarity-matching method could be sensitive to these surrounding anatomical features. This is significantly different from volume-based analysis, in which the anatomical features of the hippocampus would be contracted into one number.

4.2. Results of age-estimation

We first tested the efficacy of the MAV approach using age-dependent anatomical changes. The age estimation tested in this study may not have high clinical relevance, but as we knew the exact age, it was an ideal model with which to test the accuracy of our approach. The majority of the brain structures showed a high correlation between the estimated and actual ages, especially in the subcortical gray matter and the deep white matter ($R^2 = 0.7–0.8$). There was an overestimation of age in the pediatric population and an underestimation for the elderly population, leading to a regression slope of <1. This was likely due to a boundary problem, e.g., for a four-year-old test subject, only atlases with four-year-olds and older were available for the age estimation, thus leading to the overestimation of the age. The spatial difference in the R^2 maps and correlation slopes showed that the estimation accuracy and precision varied from structure to structure.

In majority of the anatomical areas, the atlas-voting approach had better estimation performance compared to volume correlation (Fig. 4). We attributed the better performance to the better ability of the atlas-voting to extract local anatomical features. There are two ways to interpret the regional variability of the estimation power. First, it is possible that certain anatomical regions have more age-dependent changes. Second, the employed tools may not be sensitive to changes in certain brain regions. The results in Fig. 4 could have been affected by both factors. For example, MAV achieved higher age-estimation accuracy for the subcortical gray matter and deep white matter structures, which tend to have simpler anatomical boundaries; but age estimation accuracy was not as good for the cortical gyri, as it is extremely difficult to achieve accurate boundary-to-boundary registration between atlases and targets. The atlas-voting strategy is still a field of active research and further investigation may be needed to achieve better performances.

4.3. Estimation of clinical attributes

In conventional multi-atlas brain segmentation studies, demographic and clinical information from the atlases is usually not available or is unused once satisfactory segmentation accuracy is achieved. In this study, we proposed a multi-atlas voting approach that enabled us to retrieve such information from the atlas database and use it to estimate the unknown attributes of new patients. In other words, each atlas is considered a classifier, and the opinions from multiple classifiers are rated and fused to reach a final decision. In this respect, the meaning of the multi-atlas library changes. If in segmentation accuracy is the only goal, a question like, “what is the minimum number of atlases that would be required to achieve accurate segmentation?” is meaningful, but if the multi-atlas library is considered a knowledge database from which we want to extract patient attributes, it needs to be enriched by cases with various anatomical and pathological conditions, as well as comprehensive demographic and clinical information.

The multi-atlas voting estimation of brain function (ADAS.11 score) demonstrated that the structures with the discriminating power were concentrated in the limbic structures of the medial temporal lobes, such as the amygdala, the hippocampus, the entorhinal cortex, the parahippocampal gyrus, the cingulum, and the fornix. Similar to the age estimation, the direct estimation of the functional outcome using the MAV provided demonstrated higher estimation accuracy compared to the conventional volume-functional correlation approach.

4.4. Diagnostic classification and comparison with existing studies

The diagnostic classification results in Table 2 also point out that the direct estimation of the diagnostic category or ADAS.11 scores by MAV outperformed the volume marker. The overall accuracy to differentiate the AD and NC was 0.90–0.92, based on the best discriminating structure or the optimized combination of the structures. The list of the discriminating structures and the AD-vs-NC discriminating power reported in this study are highly similar to those reported previously (Chupin et al., 2009; Cuingnet et al., 2011; Gerardin et al., 2009; Koikkalainen et al., 2011; Liu et al., 2012; Min et al., 2014), which reported AD/NC classification accuracies between 86%–92%, confirming the validity of the proposed multi-atlas voting approach. The estimation accuracy deteriorated to 0.63–0.82 for multi-class estimation (AD/MCI/NC). This was somewhat expected because the accuracy of a diagnosis based on cognitive assessment is not perfect; a definitive diagnosis of dementia cases can be achieved only by postmortem histology. In fact, even using the clinically measured ADAS.11 score, the multi-category classification accuracy was only 0.79 (Table 2). The performance of the multi-class classification by MVA was equivalent to or exceed the state-of-the-art algorithms. For example, in the computer-aided diagnosis of dementia (CADDementia) challenge (Bron et al., 2015), the best-performing algorithm yielded an accuracy of 63.0% in classifying AD/MCI/NC, among 29 competitors (accuracy ranging from 32% to 63%). It is also encouraging that the MAV-based approach achieved higher clinical correlation and classification accuracy than the conventional volume-based analysis with the same data and under the same multi-atlas framework. By adopting more advanced machine-learning methods, or even combining the structural volume or clinical data, it is possible that we can further improve the classification performance, but this is beyond the scope of the current proof-of-principle study. It should also be noted that classification is not the only goal of computer-aided diagnosis. Instead, quantitative descriptions of the regional features – the foci and vulnerable regions of a particular disease, such as the regional variations demonstrated in Figs. 5–7, are similarly important to clinical image readings. The functional attributes, such as cognitive and behavioral evaluations, estimated from the MRI (e.g., ADAS.11 score in this study) can also provide valuable characterization of the disease status besides categorical diagnosis.

The concept of our approach shared similarity with that of Coupe et al. (2012), who developed a technique known as scoring by non-local image patch estimator (SNIPE) for simultaneous segmentation and grading of structural MRI for Alzheimer's patients. Similarly, the authors determined an AD/NC grading measurement based on the anatomical similarity between the test subject and the training dataset in a few selected structures, and they concluded that the best classification accuracy can be achieved by combining volumetric and grading information from the hippocampus and entorhinal cortex. Despite the similarity in the ways that training data (in our case, atlases) are conceived, our work is also different from the previous work in several ways. First,

our approach was incorporated into a general framework of a multi-atlas method that can be potentially applied to any type of dementia without a priori knowledge about the affected structures, which is a feature highly important for clinical applications. In Coupe's approach, two structures (the hippocampus and the entorhinal cortex) that are most appropriate for estimating the likelihood of AD are pre-selected and manually defined in the "training" data. In the present study with the AD patients, the MAV method evaluated 289 structures defined in the atlases and could correctly identify the limbic structures as the most discriminating structures (Table 3). The overall accuracy for AD estimation was 0.90, replicating the results by Coupe et al. Another major difference from the previous work lies in the algorithm to measure anatomical similarity. In Coupe et al. (2012), the authors developed a voxel-based approach using the intensity difference within a local patch, centered on each voxel between test and training image; whereas, in our study, we developed a label-by-label procedure that integrated the local correlation of the boundary voxels along a label. In the MAV framework, it is possible to incorporate various types of similarly measure, and thus, it would be interesting future study to compare the different types of similarity functions.

4.5. Evaluation of anatomical patterns for disease classification

In the previous sections, the performance of the MAV approach was evaluated for each anatomical structure independently. An interesting question is whether we can evaluate the properties of a group of structures as a disease-specific anatomical feature. To test this idea, we evaluated the estimation accuracy for the clinical scores and disease categories between a single (the best discriminating structure) vs a group of structures. To select a group of structures, we tested the top 20 structures in the discriminating power ranking and the LASSO analysis. As shown in Table 2, the results using a test data set showed consistent improvement using the multi-structure anatomical features. For future clinical applications with heterogeneous pathological conditions, this could be an important research direction.

4.6. Limitations and future directions

One limitation of this study is that we tested only the MAV method in relatively well-characterized Alzheimer's disease for proof-of-concept purposes. The results of this study showed that the accuracy of this new approach was comparable to or slightly better than the conventional volume-based analyses. However, this is only the first step toward our long-term goal to evaluate the efficacy of the tools that emulate the thought process of radiologists through the CBIR approach. Our results for the region-specific estimation of functional deficits and the categorical classification in AD, without a priori knowledge about the locations of the abnormalities, demonstrated the potential to use this approach for different types of neurodegenerative diseases or for an assortment of dementia pathology. To be truly useful as a tool to support clinical evaluation of patient images, atlases with various types of pathology (e.g., AD, Lewy bodies, Parkinson's, frontotemporal dementia, vascular dementia, etc.) need to be included in the knowledge database (atlases) and the diagnostic specificity needs to be tested. Another important future direction is to incorporate non-anatomical features (demographic information, lab tests, clinical data, etc.) in the diagnostic estimator. Ultimately, the diagnostic or functional estimation are useful only if these factors can contribute to the prediction of clinically important functional deficits or treatment efficacy, which will require longitudinal data for validation. This proof-of-concept study is, thus, only the initial step in the development of clinically useful tools.

5. Conclusion

We proposed a novel MAV diagram to directly retrieve non-imaging patient attributes, in which the clinical information from multiple atlases were weighted and fused to estimate the patient's diagnosis. We demonstrated the performance of this approach in age estimation, as well as in the estimation of functional deficits and diagnostic categories in Alzheimer's disease. Although MAV is a markedly different concept than conventional volume-based anatomical analyses, our results indicated that MAV can classify patients with the comparable accuracy and detect structures that are known to be affected by AD, indicating the proof-of-principle of this new approach.

Acknowledgment

This publication was made possible by the following grants: P41EB015909 (MIM, MS), R01EB017638 (MIM), R01NS084957 (MS).

Conflict of interest

MS and MIM own "AnatomyWorks". MS is its CEO. This arrangement is being managed by the Johns Hopkins University in accordance with its conflict of interest policies.

References

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

Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de-Solorzano, C., 2009. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imaging 28, 1266–1277.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol 57, 289–300.

Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Mendez Orellana, C., Meijboom, R., Pinto, M., Meireles, J.R., Garrett, C., Bastos-Leite, A.J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cardenas-Pena, D., Alvarez-Meza, A.M., Dolph, C.V., Iftekharuddin, K.M., Eskildsen, S.F., Coupe, P., Fonov, V.S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K.R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G.M., Stoyanov, Z.V., Sorensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J.C., Niessen, W.J., Klein, S., Alzheimer's Disease Neuroimaging, I., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage 111, 562–579.

Christensen, G.E., Rabbitt, R.D., Miller, M.I., 1996. Deformable templates using large deformation kinematics. IEEE Trans. Image Process. 5, 1435–1447.

Chupin, M., Hammers, A., Liu, R.S., Colliot, O., Burdett, J., Bardinet, E., Duncan, J.S., Garnero, L., Lemieux, L., 2009. Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. NeuroImage 46, 749–761.

Collins, D.L., Holmes, C.J., Peters, T.M., Evans, A.C., 1995. Automatic 3-D model-based neuroanatomical segmentation. Hum. Brain Mapp. 3, 190–208.

Coupe, P., Eskildsen, S.F., Manjon, J.V., Fonov, V.S., Collins, D.L., Alzheimer's disease Neuroimaging, I., 2012. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. NeuroImage 59, 3736–3747.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehericy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., Alzheimer's Disease Neuroimaging, I., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56, 766–781.

Dawant, B.M., Hartmann, S.L., Thirion, J.P., Maes, F., Vandermeulen, D., Demaerel, P., 1999. Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations: part I, methodology and validation on normal subjects. IEEE Trans. Med. Imaging 18, 909–916.

Djamanakova, A., Tang, X., Li, X., Faria, A.V., Ceritoglu, C., Oishi, K., Hillis, A.E., Albert, M., Lyketsos, C., Miller, M.I., Mori, S., 2014. Tools for multiple granularity analysis of brain MRI data for individualized image analysis. NeuroImage 101, 168–176.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Gerardin, E., Chetelat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehericy, S., Garnero, L., Eustache, F., Colliot, O., Initi, A.s.D.N., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. NeuroImage 47, 1476–1486.

Gholipour, A., Kehtarnavaz, N., Briggs, R., Devous, M., Gopinath, K., 2007. Brain functional localization: a survey of image registration techniques. IEEE Trans. Med. Imaging 26, 427–451.

Grenander, U., Miller, M.I., 1998. Computational anatomy: an emerging discipline. Q. Appl. Math. 56, 617–694.

Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115–126.

Heckemann, R.A., Hammers, A., Rueckert, D., Aviv, R.I., Harvey, C.J., Hajnal, J.V., 2008. Automatic volumetry on MR brain images can support diagnostic decision making. BMC Med. Imaging 8, 9.
Hsu, W., Taira, R.K., El-Saden, S., Kangarloo, H., Bui, A.A., 2012. Context-based electronic health record: toward patient specific healthcare. IEEE Trans. Inf. Technol. Biomed. 16, 228–234.
Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M.A., van Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusion—application to cardiac and aortic segmentation in CT scans. IEEE Trans. Med. Imaging 28, 1000–1010.
Jia, H., Yap, P.T., Shen, D., 2012. Iterative multi-atlas-based multi-image segmentation with tree-based registration. NeuroImage 59, 422–430.
Joshi, S., Davis, B., Jomier, M., Gerig, G., 2004. Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23 (Suppl 1), S151–S160.
Klein, A., Mensh, B., Ghosh, S., Tourville, J., Hirsch, J., 2005. Mindboggle: automated brain labeling with multiple atlases. BMC Med. Imaging 5, 7.
Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46, 786–802.
Kloppel, S., Stonnington, C.M., Barnes, J., Chen, F., Chu, C., Good, C.D., Mader, I., Mitchell, L.A., Patel, A.C., Roberts, C.C., Fox, N.C., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S., 2008. Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. Brain 131, 2969–2974.
Koikkalainen, J., Lotjonen, J., Thurfjell, L., Rueckert, D., Waldemar, G., Soininen, H., Alzheimer's Disease Neuroimaging, I., 2011. Multi-template tensor-based morphometry: application to analysis of Alzheimer's disease. NeuroImage 56, 1134–1144.
Langerak, T.R., van der Heide, U.A., Kotte, A.N.T.J., Viergever, M.A., van Vulpen, M., Pluim, J.P.W., 2010. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans. Med. Imaging 29, 2000–2008.
Liang, Z., He, X., Ceritoglu, C., Tang, X., Li, Y., Kutten, K.S., Oishi, K., Miller, M.I., Mori, S., Faria, A.V., 2015. Evaluation of cross-protocol stability of a fully automated brain multi-atlas parcellation tool. PLoS One 10, e0133533.
Liu, M., Zhang, D., Shen, D., Alzheimer's Disease Neuroimaging, I., 2012. Ensemble sparse classification of Alzheimer's disease. NeuroImage 60, 1106–1116.
Llano, D.A., Laforet, G., Devanarayan, V., Initia, A.D.N., 2011. Derivation of a new ADAS-cog composite using tree-based multivariate analysis prediction of conversion from mild cognitive impairment to Alzheimer disease. Alzheimer Dis. Assoc. Disord. 25, 73–84.
Lotjonen, J., Koikkalainen, J., Thurfjell, L., Rueckert, D., 2009. Atlas-Based Registration Parameters in Segmenting Sub-Cortical Regions From Brain Mri-Images. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Vols 1 and 2. pp. 21–24.
Lotjonen, J.M.P., Wolz, R., Koikkalainen, J.R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D., Initi, A.s.D.N., 2010. Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage 49, 2352–2365.
Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality Image Registration by Maximization of Mutual Information. Medical Imaging, IEEE Transactions on 16. pp. 187–198.
Miller, M.I., Christensen, G.E., Amit, Y., Grenander, U., 1993. Mathematical textbook of deformable neuroanatomies. Proc. Natl. Acad. Sci. U. S. A. 90, 11944–11948.
Miller, M.I., Mori, S., Tang, X., Tward, D., Zhang, Y., 2015. Bayesian multiple atlas deformable templates. In: Toga, A.W. (Ed.), Brain Mapping: An Encyclopedic Reference. Academic Press: Elsevier, pp. 401–415.
Min, R., Wu, G., Cheng, J., Wang, Q., Shen, D., Alzheimer's Disease Neuroimaging, I., 2014. Multi-atlas based representations for Alzheimer's disease diagnosis. Hum. Brain Mapp. 35, 5052–5070.
Mori, S., Oishi, K., Faria, A.V., Miller, M.I., 2013. Atlas-based neuroinformatics via MRI: harnessing information from past clinical cases and quantitative image analysis for patient care. Annu. Rev. Biomed. Eng. 15 (15), 71–92.
Muller, H., Michoux, N., Bandon, D., Geissbuhler, A., 2004. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int. J. Med. Inform. 73, 1–23.
Rohlfing, T., Maurer Jr., C.R., 2007. Shape-based averaging. IEEE Trans. Image Process. 16, 153–161.
Rohlfing, T., Brandt, R., Menzel, R., Maurer, C.R., 2004. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocol microscopy images of bee brains. NeuroImage 21, 1428–1442.
Sabuncu, M.R., Yeo, B.T.T., Van Leemput, K., Fischl, B., Golland, P., 2010. A generative model for image segmentation based on label fusion. IEEE Trans. Med. Imaging 29, 1714–1729.
Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R., 2000. Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 22, 1349–1380.
Tang, L.H.Y., Hanka, R., Ip, H.H.S., 1999. A review of intelligent content-based indexing and browsing of medical images. Health Informatics J. 5, 40–49.
Tang, X., Oishi, K., Faria, A.V., Hillis, A.E., Albert, M.S., Mori, S., Miller, M.I., 2013. Bayesian parameter estimation and segmentation in the multi-atlas random orbit model. PLoS One 8, e65591.
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B Methodol 58, 267–288.
Tustison, N.J., Cook, P.A., Klein, A., Song, G., Das, S.R., Duda, J.T., Kandel, B.M., van Strien, N., Stone, J.R., Gee, J.C., Avants, B.B., 2014. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. NeuroImage 99, 166–179.
van Rikxoort, E.M., Isgum, I., Arzhaeva, Y., Staring, M., Klein, S., Viergever, M.A., Pluim, J.P.W., van Ginneken, B., 2010. Adaptive local multi-atlas segmentation: application to the heart and the caudate nucleus. Med. Image Anal. 14, 39–49.
Wang, H., Suh, J.W., Das, S.R., Pluta, J., Craige, C., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. 35, 611–623.
Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23, 903–921.
Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C.S., Aizenstein, H.J., 2007. Optimum template selection for atlas-based segmentation. NeuroImage 34, 1612–1618.
Wu, D., Ma, T., Ceritoglu, C., Li, Y., Chotiyanonta, J., Hou, Z., Hsu, J., Xu, X., Brown, T., Miller, M.I., Mori, S., 2015. Resource atlases for multi-atlas brain segmentations with multiple ontology levels based on T1-weighted MRI. NeuroImage 125, 120–130.