Feature-based vs. intensity-based brain image registration: voxel level and structure level performance evaluation

Leonid Teverovskiy, Owen Carmichael,
Howard Aizenstein, Nicole Lazar, Yanxi Liu

November 2006
CMU-ML-06-118

Carnegie Mellon

Feature-based vs. intensity-based brain image registration: voxel level and structure level performance evaluation

Leonid A. Teverovskiy[1], **Owen T. Carmichael**[2],
Howard J. Aizenstein[3], **Nicole A. Lazar**[4] **and Yanxi Liu**[3,5,6]

November 2006
CMU-ML-06-118

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Department of Neurology, University of California, Davis, CA, USA
[3]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA.
[4]Department of Statistics, University of Georgia, Athens, GA, USA
[5]Department of Computer Science and Engineering, Penn State University, University Park, PA, USA
[6]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA

This research is supported by the grant P50-AG05133 from National Institute of Health and PA tobacco grant 4100027294.

Keywords: deformable registration evaluation, mutual information, false discovery rate

Abstract

The power and validity of voxel based and tensor based morphometry methods depend on the accuracy of the brain image registration algorithms they employ. We propose a mutual information based quantitative evaluation method to compare the performance of two publicly available deformable registration packages: HAMMER and algorithms in the ITK package (FEM-Demons) The advantage of our approach is that registration algorithms are quantitatively compared at both global and local levels, thus enabling our method to pinpoint areas of the brain where one algorithm performs significantly better or worse than the others. The brain image dataset used for evaluation consists of a total of 59 images: 20 MR images of Alzheimer's (AD) patients, 19 MR images of people with mild cognitive impairment (MCI) and 20 MR images of normal (CTL) subjects. Global and localized mutual information scores are used to evaluate the quality of registration, and paired t-tests are used to determine the statistical significance of registration quality differences between the methods at three levels: global, voxel-wise and anatomical structures. We threshold the resulting p-value maps using a false discovery rate control method in order to correct for multiple comparisons. Our results show that both HAMMER and FEM-Demons algorithms do significantly better than an affine registration algorithm, FLIRT, at all three levels for all three subject groups. Comparison between the HAMMER and FEM-Demons algorithms shows that at the global level there is no significant difference in performance between the two algorithms on controls, and FEM-Demons outperforms HAMMER on Alzheimers patients (p-value 0.0416) and MCI patients (p-value 0.0055). At the local and anatomical levels, FEM-Demons and HAMMER dominate each other on different brain regions. Our results indicate that the choice between the HAMMER and the FEM-Demons algorithms should depend on the region of interest of a study.

1 Introduction

Deformable registration algorithms are at the center of voxel-based morphometry methods(VBM)[10], tensor field based morphometry methods (TBM) [16, 5, 31] and many automated segmentation methods [15, 27]. The strengths and limitations of VBM, TBM and atlas based segmentation methods depend on the strength and limitations of the deformable registration algorithms they employ, and therefore it is crucial to explicitly state the accuracy of the registration when reporting scientific findings. We propose a mutual information based methodology for evaluating and comparing registration algorithms. Our goal is to evaluate one of the necessary conditions for a good deformable registration: visual similarity between registered and reference images. A popular and well-studied measure of image similarity is the mutual information [4, 33]. In this paper we employ global and localized mutual information between the template and registered images in order to quantitatively evaluate the performance of a registration algorithm. Our framework is fully automated and is applicable in situations where manual segmentation of anatomical structures is not available. Paired T-tests with multiple comparison correction are used to find areas of the brain where the performance of one algorithm is significantly better than that of another. We apply this methodology to two fully deformable registration algorithms and one affine registration algorithm: finite element (FEM) based registration followed by demons registration algorithms available as part of the Insight Toolkit (ITK) [13, 29, 30, 8], hierarchical attribute matching mechanism for elastic registration (HAMMER) algorithm [25], and FMRIB's Linear Image Registration Tool (FLIRT) [14]. The registration algorithms we use in ITK are intensity-based and fully deformable; HAMMER is a feature-based fully deformable registration algorithm, and FLIRT is an intensity-based affine registration algorithm. As an affine registration algorithm, FLIRT is constrained to transformations with 12 degrees of freedom, which limits its registration ability compared to the fully deformable HAMMER and FEM-Demons algorithms. Thus, comparing FLIRT with FEM-Demons and FLIRT with HAMMER helps us validate our evaluation method. Then, we apply our method to compare HAMMER and FEM-Demons registration algorithms in order to highlight their areas of strength. The paper is organized as follows. Section 2 introduces relevant work in the area of registration comparison. Section 3 describes data that was used for the experiments. The evaluation procedure is described in detail in Section 4. Section 5 presents registration comparison results at global, voxel-wise and anatomical structure levels. Section 6 contains the discussion of the results, followed by a summary in Section 7.

2 Related Work

There are several existing methods for the evaluation of registration algorithms. The approach which evaluates intermodality rigid body registration algorithms in the retrospective image registration evaluation project [6] relies on landmarks in the template and test images. The quality of registration is measured by the Euclidean distance between the corresponding landmarks in the template and registered test images. The smaller is this distance, the better is the performance of the registration algorithm.

Another very popular approach is to evaluate a registration algorithm through its performance

in an automated segmentation task. A human expert delineates certain anatomical structures on the template and test images by hand. After the registration algorithm is applied to register the test image to the template, an overlap percentage between the delineated anatomical structures of the template and test images is computed. Better image registration corresponds to greater overlap percentage [2, 27, 15, 37]. Obtaining ground truth for these methods is an extremely labor-intensive task, and developing a tracing protocol for segmentation is very time consuming and complex. Also, segmentation results are rater-dependent. Usually, only a few anatomical structures are segmented, and a registration algorithm is evaluated only on these structures.

Another technique used for quantitative evaluation of the registration algorithms involves using simulated deformations. An image is deformed using an artificially created deformation field, and then the deformed image is registered to the original image. The difference between the known artificial deformation field and the field estimated by a registration algorithm serves as a basis for evaluating the algorithm [34]. In this case, there is no additional work required to obtain the ground truth. However, it should be noted that registering an image to a deformed version of itself removes inter-subject and inter-scan variability as obstacles to accurate registration.

A recently developed method [23] evaluates the performance of the registration algorithms on a set of images by measuring generalisation and specificity of the brain appearance models estimated based on the registration results. The more similar the distribution of the images generated by the appearance model is to the distribution of the images that are used to estimate the model, the more accurate the registration algorithm is. However, the evaluation results obtained by this method depend on the functional form of the appearance model, the number of its parameters and how well these parameters can be estimated.

In contrast, we propose an evaluation scheme that does not require ground truth or appearance model estimation. Our method quantitatively evaluates one necessary condition for good deformable registration of images: visual similarity. In addition, we compare the performance of deformable registration algorithms on images from different perspectives: as a whole, at voxel-wise level and at the level of anatomical structures.

3 Data

Our test dataset consists of structural MR images of 59 subjects. These images were acquired on GE 1.5T Signa scanner between 1999 and 2004 at the University of Pittsburgh Alzheimer's Research Center. The spoiled gradient-recalled (SPGR) volumetric T1-weighted pulse sequence is used with the following parameters optimized for maximal contrast among gray matter, white matter, and CSF: TE = 5ms, TR = 25ms, flip angle = 40, NEX = 1, slice thickness = 1.5 mm/0 mm interslice. Based on a series of clinical and neurophysiological tests, the 59 subjects were diagnosed into 20 Alzheimer's (AD) patients, 19 patients with mild cognitive impairment (MCI), and 20 controls. MCI and AD cause brain changes that induce a high degree of variability in brain structure, thus making registration more difficult. We removed skulls from all the images in our dataset using the BET tool [28].

We use the anatomical automatic labeling (AAL) digital atlas [12, 32] distributed as a part of the MRIcro package as a reference image. The atlas is an MR image in MNI space [20] accompanied

by the labeling of 116 anatomical structures.

4 Method

4.1 Registration Algorithms

We use two fully deformable registration methods and an affine registration method: finite element based registration followed by demons registration available as part of Insight Toolkit library [13, 29, 30, 8], hierarchical attribute matching mechanism for elastic registration (HAMMER) developed by Shen et al [25], and FMRIB's Linear Image Registration Tool (FLIRT). We use the respective default parameters for all methods in our evaluation.

4.1.1 FEM-Demons

The Insight Toolkit provides a finite element based registration method [13, 8, 7] and demons registration algorithm [13, 29, 30]. In the FEM registration algorithm, an image is modeled as a collection of elastic structural elements. The mean sum of squared differences of image intensities determines the external forces that act on the elastic elements. The deformation field between images is calculated based on the physical properties of the structural elements and the external forces. This procedure helps avoid local maxima and guarantees that the resulting deformation field is smooth. The FEM registration algorithm is an iterative approach that employs a multiresolution scheme. Once a smooth deformation field between images is found using finite element based registration, we use the demons algorithm to estimate the deformation field more precisely. The demons registration algorithm treats an image as a set of iso-intensity contours. Displacement at each voxel is determined according to optical flow between the reference image and the image that is being registered to it.

4.1.2 HAMMER

The hierarchical attribute matching mechanism for elastic registration uses a very different image similarity measure. It computes a set of predefined attributes to find correspondences between the voxels of the reference image and an image of a subject. HAMMER requires that images are segmented into white matter, gray matter and CSF prior to registration. *Fast* tool [35] from the Oxford Center for Functional Magnetic Resonance Imaging of the Brain software library (FSL) is used to perform the segmentation. HAMMER computes an attribute vector for each voxel in the image. The attribute vector contains the intensity of the voxel, the edge type of the voxel and the geometric moment invariants about this voxel for each tissue type [25]. The correspondences between voxels are determined based on the similarities of their attribute vectors. The deformation field between images is estimated and refined over several iterations. During each iteration the deformation field between images is computed based on the correspondences between the driving voxels, i.e. voxels

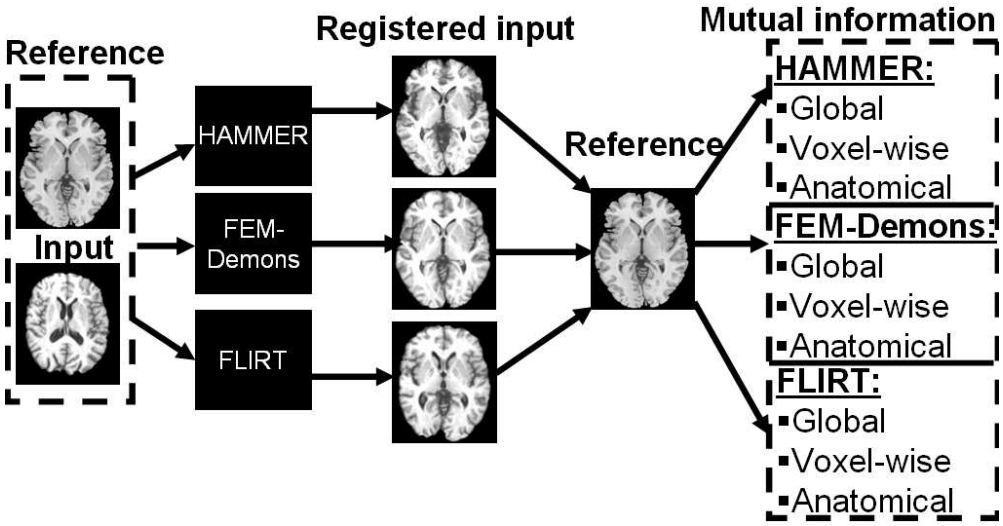


Figure 1: HAMMER, FEM-Demons and FLIRT algorithms are used to register an input image to the reference image. Registration quality is evaluated using global and localized mutual information between deformed input image and the reference image.

with the most unique attribute vectors. The number of driving voxels is increased from iteration to iteration. HAMMER employs a multiresolution scheme to improve efficiency and avoid local maxima.

4.1.3 FLIRT

FMRIB's Linear Image Registration Tool is a robust affine registration algorithm. It utilizes a mutliresolution scheme and a combination of the Powell optimization method [22] with an exhaustive search over rotation angles. This algorithm is widely used for affine registration, and is shown to perform as well as or better than other popular affine registration algorithms, including AIR and SPM [36]. We use the default correlation ratio similarity metric for our experiments.

4.2 Registration comparison

We use global and localized mutual information to evaluate the registration algorithms. Mutual information is a well-studied and widely used measure of similarity between images [4, 33]. Similar images have high mutual information scores because they explain each other well. We use the following definition of mutual information:

$$I(A, B) = H(A) + H(B) - H(A, B)$$

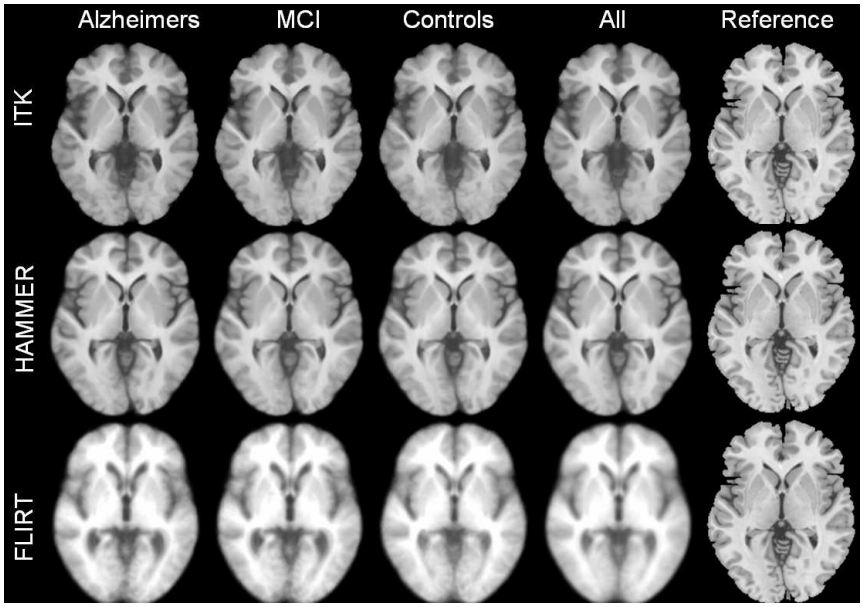


Figure 2: Corresponding slices of averaged registered images by HAMMER, FEM-Demons and FLIRT. Averages per class as well as average per dataset are shown. Last column contains corresponding slice of the reference brain.

where $I(A, B)$ is mutual information between images A and B,

$$H(A) = \sum_{a \in A} p(a) \log p(a)$$

and

$$H(B) = \sum_{b \in B} p(b) \log p(b)$$

are the Shannon entropies of gray level value distributions of images A and B, and

$$H(A, B) = \sum_{a \in A, b \in B} p(a, b) \log p(a, b)$$

is the Shannon entropy [24] of the joint distribution of gray level values of the images A and B. a and b denote gray level values of the images A and B respectively.

In addition to computing mutual information between each registered image in our dataset and the *AAL* template, we also compute localized mutual information between the corresponding 3x3x3, 7x7x7 and 11x11x11 voxel neighborhoods around each voxel in the reference and registered images. In order to assess the quality of registration for various anatomical structures of the brain, we compute mutual information for 116 anatomical structures, as defined by the Anatomical Automatic Labeling (AAL) atlas [32].

The three different registration methods provide us with three different sets of registered images. We compute global, voxel-wise and anatomical mutual information scores between each

Table 1: Comparison between the registration methods on the global scale. Subscripts *H* or *I* near pairwise t-test statistics indicate that mutual information scores for HAMMER or FEM-Demons are higher, respectively. * indicates statistical significance at 0.05 level

Registration Methods	Subject group		
	Alzheimers	MCI	Controls
FEM-Demons vs HAMMER	2.19^*_I	3.15^*_I	0.50_I
FEM-Demons vs FLIRT	35.24^*_I	37.42^*_I	35.78^*_I
HAMMER vs FLIRT	48.19^*_H	42.18^*_H	50.79^*_H

pair of the three registered image sets. A paired T-test is used to determine whether the difference in global mutual information for the pairs of registration methods is statistically significant. We also use paired T-tests with multiple comparison correction to find voxels and anatomical structures where voxel-wise and anatomical structure registration scores differ significantly. It is worth noting that none of the three algorithms uses mutual information as a similarity measure in their respective registration processes, and so using mutual information for comparing them is not unfairly advantageous to any of the algorithms.

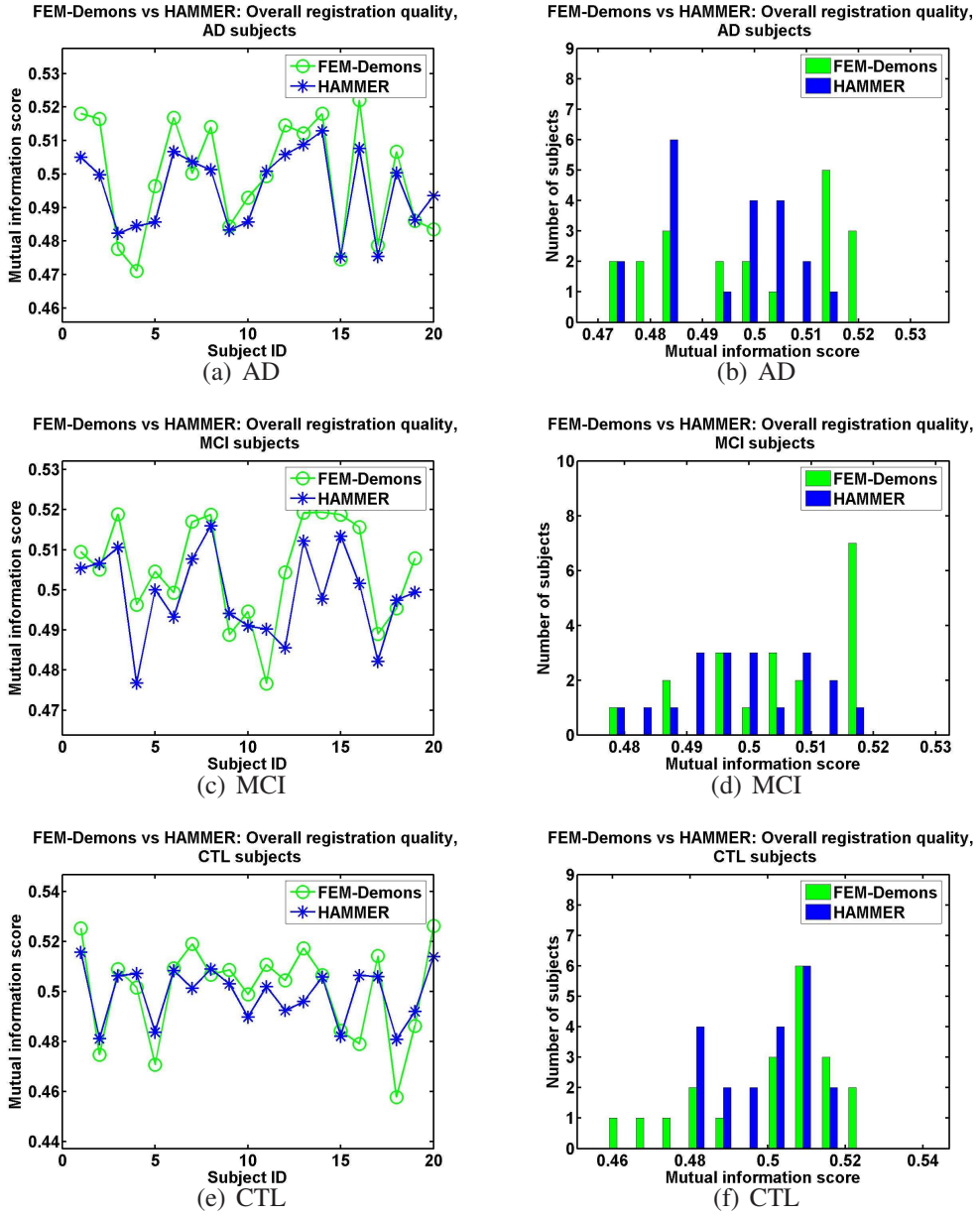
5 Comparison results

5.1 Registration results

Each image in the dataset is registered to the reference image by HAMMER, FEM-Demons and FLIRT. Deformation fields obtained using each of the registration methods are then used to transform each original image to a *registered* image. The registration procedure is illustrated in Figure 1. Averaged registered images are shown in Figure 2. Mutual information is used to quantify the similarity between each registered image and the reference image.

5.2 Global comparison

The global mutual information score is the mutual information computed between the entire reference and registered images. It reveals (see Figure 3 and Table 1)that while there is no significant difference at 95% level between HAMMER and FEM-Demons for the controls (p-value=0.6218), FEM-Demons performs significantly better on Alzheimers patients(p-value=0.0416) and MCI patients (p-value=0.0055). As expected, both HAMMER and FEM-Demons perform significantly better than FLIRT (see Figures 7, 8 in the Appendix).



(a) AD (b) AD

(c) MCI (d) MCI

(e) CTL (f) CTL

Figure 3: Global mutual information registration scores for HAMMER and FEM-Demons. a, b - Alzheimers patients; c, d - MCI patients; e, f - controls. Figures b, d, f show histograms of the MI score distribution between the two methods.

5.3 Multiple Comparison Correction

There are two groups of methods for performing a multiple comparison correction. The first group controls the family-wise error[11, 21]. Methods following this approach adjust p-value thresholds of the individual tests so that the probability of type I error occurring in *any* of the tests in the family is no greater than a given significance level. Bonferroni correction is the most popular of these methods. The p-value thresholds adjusted in this way are generally very conservative because they have to ensure that the occurrence of even a single false positive in the entire set of tests is unlikely. Controlling family-wise error is a very intuitive thing to do in a situation where the conclusion for the entire family can be drawn from a single positive test.

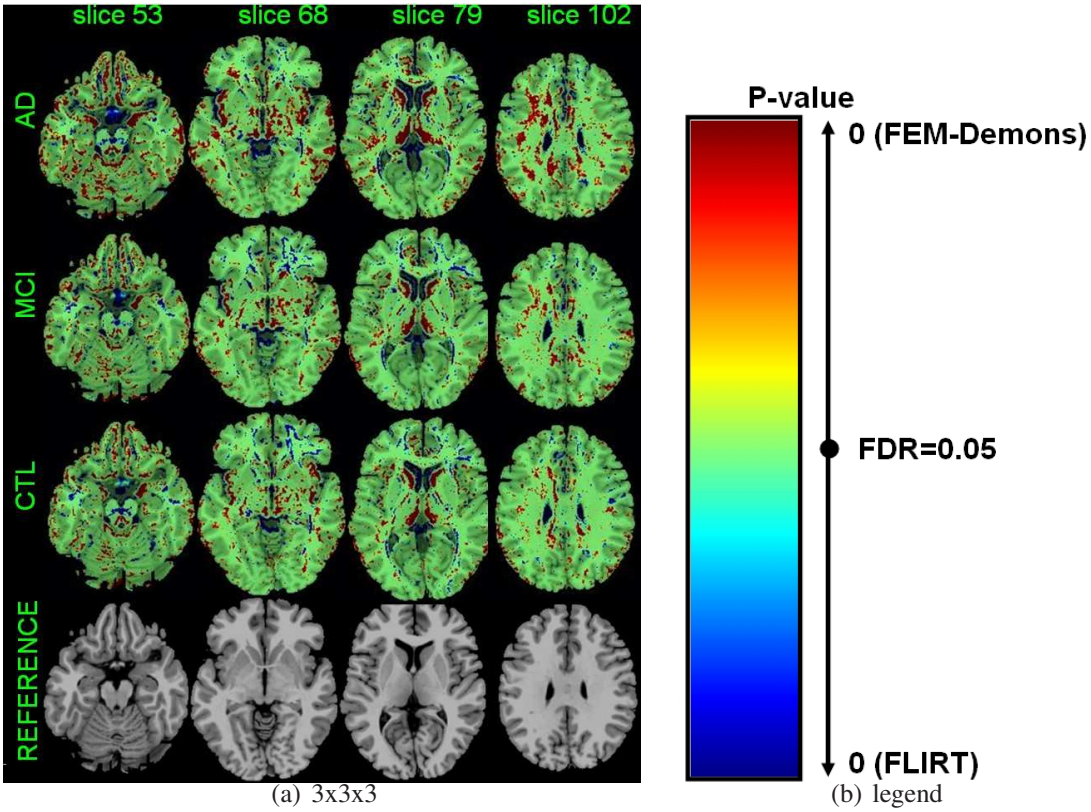
However, the conservativeness of these methods comes at the expense of their power. The second group of methods adjusts the p-value threshold of the individual tests so that the expected ratio of false positives to the total number of positives is no greater than a given false discovery rate bound [9, 1]. These methods are applicable where the focus of an analysis is to provide a descriptive statistic about a family of tests, but not to use individual tests to draw conclusions for the family. In other words, adjusting false discovery rate is particularly suitable for the situations where a researcher is willing to tolerate a small proportion of false positives in exchange for a greater test power (i.e. smaller probability of a false negative).

The goal of our analysis is to find brain regions where one method performs statistically better than the other. It is important for our analysis not to erroneously declare voxels where the two registration methods perform comparably as significant voxels. However, it is equally important to successfully detect voxels where the performance of the two registration algorithms is statistically different. Therefore, we are willing to allow a small percentage of false positives in exchange for the greater power of individual tests. Methods that control false discovery rate fit our needs perfectly, and we use them to perform multiple comparison correction in our analysis. In order to demonstrate that our results of the deformable registration algorithm comparison are consistent for various levels of false discovery rate, we use rates of 1%, 5% and 10%. Setting false discovery rate at 1% yields a p-value threshold of about 0.002 depending on the pair of algorithms being compared, neighborhood size and subject group, while false discovery rates of 5% and 10% correspond to p-value thresholds of about 0.02 and 0.05 respectively.

5.4 Voxel-wise comparison

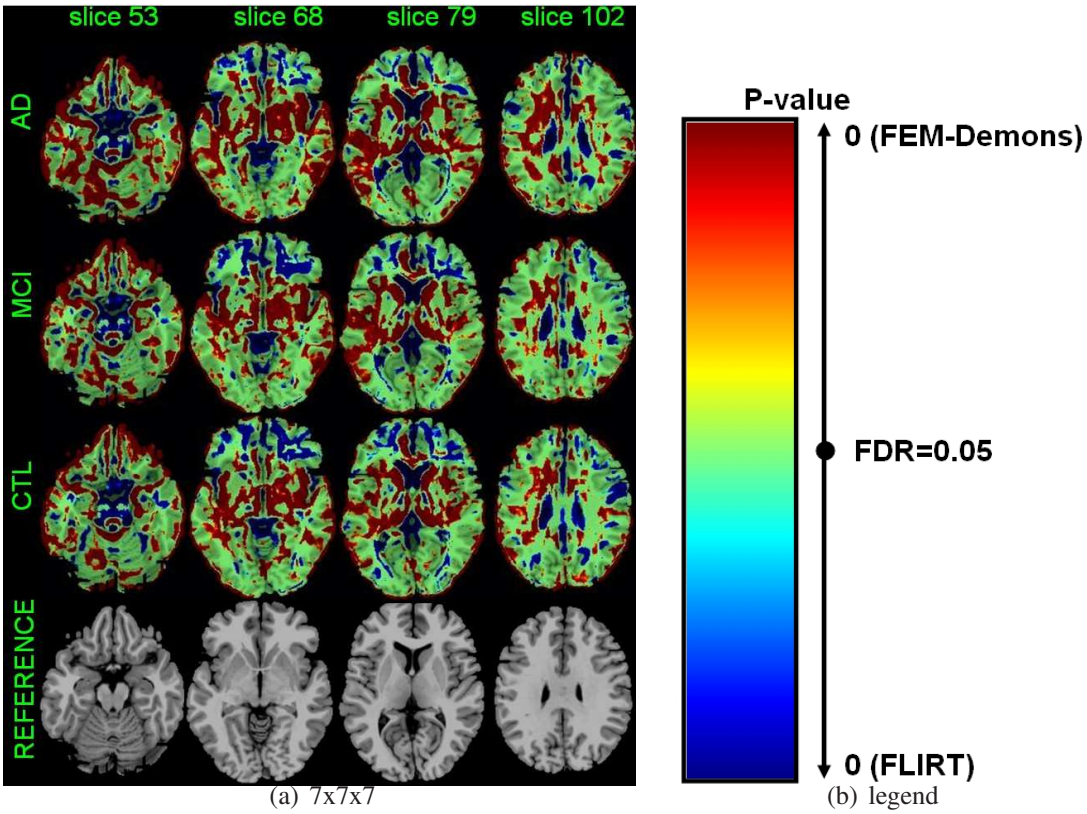
Voxel-wise comparison between the registration algorithms indicates that there are roughly twice as many voxels at which FEM-Demons significantly outperforms HAMMER than voxels where HAMMER outperforms FEM-Demons. Table 2 contains comparison results for each subject group and voxel neighborhood at three FDR levels: 0.01, 0.05 and 0.1.

P-value maps based on the localized mutual information for each subject group and neighborhood sizes of 3x3x3, 7x7x7 and 11x11x11 voxels are illustrated in Figures 4, 5, 6. Both HAMMER and FEM-Demons outperform FLIRT on over 90% of voxels. The corresponding tables and figures for comparison with FLIRT are presented in the Appendix. From these figures one can observe the tradeoff between the neighborhood size and the consistency of localized mutual information scores. The smaller is the neighborhood, the greater is the locality of the mutual information. On



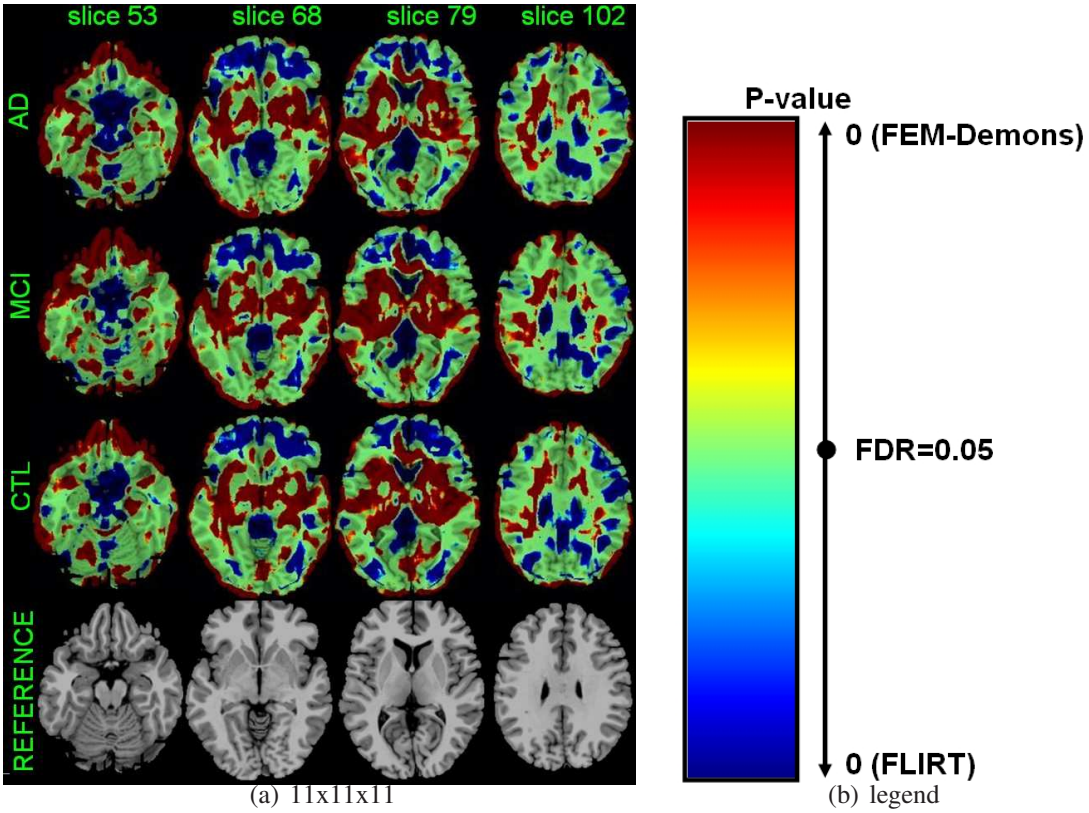
(a) 3x3x3 (b) legend

Figure 4: P-values from a paired T-test for every voxel. Red indicates voxels where MI scores for FEM-Demons registration are significantly higher than that of HAMMER; blue - voxels where HAMMER scores are higher. Intensity of colors indicates statistical significance. (a) p-values for sample slices for 3x3x3 neighborhood; (b) shows the color map)



(a) 7x7x7 (b) legend

Figure 5: P-values from a paired T-test for every voxel. Red indicates voxels where MI scores for FEM-Demons registration are significantly higher than that of HAMMER; blue - voxels where HAMMER scores are higher. Intensity of colors indicates statistical significance. (a) p-values for sample slices for 7x7x7 neighborhood; (b) shows the color map)



(a) 11x11x11 (b) legend

Figure 6: P-values from a paired T-test for every voxel. Red indicates voxels where MI scores for FEM-Demons registration are significantly higher than that of HAMMER; blue - voxels where HAMMER scores are higher. Intensity of colors indicates statistical significance. (a) p-values for sample slices for 11x11x11 neighborhood; (b) shows the color map)

the other hand, smaller neighborhoods do not provide enough samples to accurately estimate joint voxel intensity distribution required to compute mutual information. From our experiments we observed that the results based on 7x7x7 and 11x11x11 voxel neighborhoods are more consistent with one another than with 3x3x3 neighborhood, leading us to believe that at least 7x7x7 voxel or larger neighborhoods should be used for computing mutual information at the voxel-wise level.

5.5 Anatomical structure comparison

Comparison between FEM-Demons, HAMMER and FLIRT on anatomical structures is done for 116 structures using the AAL atlas [32]. The mutual information score is computed for each structure and every registered image in the dataset and the corresponding structure in the *AAL* brain. Locations of the anatomical structures are determined according to the AAL labelling. As in the case with voxel-wise comparisons, we set FDR level at 1%, 5% and 10%, which corresponds to p-value thresholds of about 0.004, 0.020 and 0.045 respectively, depending on the subject group and the pair of methods under consideration. The result summary for each subject group is presented in the Table 3, and Tables 9, 10 of the Appendix. FEM-Demons and HAMMER produce significantly better results than FLIRT did on almost all anatomical structures we considered. The difference in performance between FEM-Demons and HAMMER is the greatest for MCI patients, where FEM-Demons outperformed HAMMER on 31 anatomical structures, while HAMMER outperformed FEM-Demons on 15. Table 3 shows the comparison results at the anatomical structure level for all subject groups at the FDR levels of 0.01, 0.05 and 0.1. Tables 4, 5 and 6 show anatomical structures where FEM-Demons outperforms HAMMER and structures where HAMMER outperforms FEM-Demons.

6 Discussion

Our evaluation results confirm our expectations that both deformable registration algorithms (HAMMER and FEM-Demons) significantly outperform the affine registration algorithm (FLIRT) at all three levels: global, voxel-wise and anatomical structures. This result is not surprising, since, unlike the fully deformable registration algorithms, FLIRT is limited to only 12 degrees of freedom. This result serves as a sanity check for our evaluation method.

Our comparison results on the two fully deformable registration algorithms show that both HAMMER and FEM-Demons have areas of the brain where they outperform the other. Although FEM-Demons outperforms HAMMER on roughly twice as many voxels, HAMMER outperforms FEM-Demons on structures with well-defined boundaries, for example, the ventricles and the posterior cingulate. This phenomenon could be explained by the fact that HAMMER employs image segmentation into white matter, gray matter and CSF during registration process, while the algorithms currently included in FEM-Demons do not. Similar reasoning could explain why FEM-Demons algorithm does better in homogeneous regions, like the thalamus and the caudate: FEM-Demons uses information from intensity variations within these regions, while HAMMER, since it only uses segmented images, does not.

At all three levels, the difference in performance between HAMMER and FEM-Demons is subject group dependent (see Tables 1, 2, 7, 8, 3, 9, 10). For example, Table 1 shows that at the global level FEM-Demons and HAMMER do comparably well on MR images of normal subjects, while FEM-Demons outperforms HAMMER on Alzheimers patients (p-value is 0.0416), and especially on MCI patients (p-value 0.006). A

Table 2: FEM-Demons vs HAMMER: percentage of voxels where mutual information scores for one method are significantly higher than for the other.

3x3x3			
Registration Method	Alzheimers	MCI	Controls
FEM-Demons			
FDR=0.01	7.91%	5.26%	5.43%
FDR=0.05	17.74%	13.28%	13.28%
FDR=0.10	30.21%	24.58%	24.24%
HAMMER			
FDR=0.01	2.76%	2.95%	3.09%
FDR=0.05	6.31%	6.95%	7.31%
FDR=0.10	11.70%	13.29%	14.02%
7x7x7			
Registration Method	Alzheimers	MCI	Controls
FEM-Demons			
FDR=0.01	21.95%	18.57%	17.97%
FDR=0.05	32.20%	28.27%	27.44%
FDR=0.10	42.07%	37.93%	36.93%
HAMMER			
FDR=0.01	6.83%	6.42%	6.58%
FDR=0.05	11.62%	11.33%	11.97%
FDR=0.10	17.25%	17.61%	18.63%
11x11x11			
Registration Method	Alzheimers	MCI	Controls
FEM-Demons			
FDR=0.01	21.07%	21.76%	19.43%
FDR=0.05	29.47%	30.52%	27.84%
FDR=0.10	37.62%	39.02%	35.88%
HAMMER			
FDR=0.01	10.22%	8.32%	9.37%
FDR=0.05	16.61%	14.26%	16.22%
FDR=0.10	23.44%	20.89%	23.54%

Table 3: FEM-Demons vs HAMMER: number and percentage of anatomical regions where mutual information scores for one method are significantly higher than for the other. Total of 116 regions were considered, as segmented on the AAL atlas.

Registration Method	Alzheimers	MCI	Controls
FEM-Demons			
FDR=0.01	13 (11.21%)	18 (15.52%)	11 (9.48%)
FDR=0.05	26 (22.41%)	31 (26.72%)	16 (13.79%)
FDR=0.10	34 (29.31%)	33 (28.45%)	22 (18.97%)
HAMMER			
FDR=0.01	14 (12.07%)	13 (11.21%)	13 (11.21%)
FDR=0.05	21 (18.10%)	15 (12.93%)	19 (16.38%)
FDR=0.10	23 (19.83%)	18 (15.52%)	24 (20.69%)

Table 4: FEM-Demons vs HAMMER, AD subjects: List of anatomical regions where mutual information scores for one method are significantly (FDR=0.05) higher than for the other.

FEM-Demons	Cerebelum_7b_L, Cerebelum_7b_R, Cerebelum_Crus2_R, Frontal_Sup_Orb_L, Frontal_Sup_Orb_R, Frontal_Sup_R, Heschl_L, Occipital_Mid_R, Parietal_Inf_L, Parietal_Inf_R, Postcentral_L, Supp_Motor_Area_L, Supp_Motor_Area_R, SupraMarginal_L, Temporal_Inf_L, Temporal_Inf_R, Temporal_Mid_L, Temporal_Pole_Mid_L, Temporal_Pole_Sup_L, Temporal_Sup_L, Temporal_Sup_R
HAMMER	Cerebelum_3_L, Cerebelum_3_R, Cerebelum_9_L, Cerebelum_9_R, Cingulum_Mid_L, Cingulum_Post_L, Cingulum_Post_R, Cuneus_R, Frontal_Inf_Oper_L, Frontal_Mid_Orb_L, Precentral_R, Precuneus_R, Vermis_10, Vermis_1_2, Vermis_3, Vermis_4_5, Vermis_6

Table 5: FEM-Demons vs HAMMER, MCI subjects: List of anatomical regions where mutual information scores for one method are significantly (FDR=0.05) higher than for the other.

FEM-Demons	Amygdala_R, Angular_L, Cerebelum_7b_L, Cerebelum_7b_R, Cerebelum_8_R, Frontal_Sup_Medial_L, Frontal_Sup_Medial_R, Frontal_Sup_Orb_L, Frontal_Sup_Orb_R, Heschl_L, Heschl_R, Insula_L, Insula_R, Parietal_Inf_L, Postcentral_L, Putamen_L, Rectus_L, Rolandic_Oper_L, Rolandic_Oper_R, SupraMarginal_L, Temporal_Inf_R, Temporal_Mid_L, Temporal_Pole_Mid_L, Temporal_Pole_Sup_L, Temporal_Sup_L, Thalamus_R
HAMMER	Cerebelum_3_L, Cerebelum_3_R, Cingulum_Mid_L, Cingulum_Post_L, Cingulum_Post_R, Frontal_Inf_Oper_L, Occipital_Sup_R, Precuneus_R, Vermis_10, Vermis_1_2, Vermis_3, Vermis_4_5, Vermis_6

Table 6: FEM-Demons vs HAMMER, CTL subjects: List of anatomical regions where mutual information scores for one method are significantly (FDR=0.05) higher than for the other.

FEM-Demons	Cerebelum_10_L, Cerebelum_7b_L, Cerebelum_7b_R, Cerebelum_8_L, Cerebelum_Crus2_R, Frontal_Sup_Orb_L, Frontal_Sup_Orb_R, Heschl_L, Olfactory_L, Postcentral_L, Rectus_L, SupraMarginal_L, Temporal_Inf_R, Temporal_Pole_Mid_L, Temporal_Sup_L
HAMMER	Cerebelum_3_L, Cingulum_Mid_L, Cingulum_Post_L, Cingulum_Post_R, Cuneus_L, Cuneus_R, Occipital_Mid_L, Occipital_Sup_R, Paracentral_Lobule_R, Precuneus_R, Vermis_10, Vermis_1_2, Vermis_3, Vermis_4_5, Vermis_6, Vermis_9

possible reason for these results is that automatic MR image segmentation, which HAMMER relies on, is more accurate for controls than for patients.

Our results show that FLIRT outperforms FEM-Demons and HAMMER on 9% to 14% of the voxels when 3x3x3 neighborhood is used. The reason for having voxels where FLIRT has higher mutual information scores than the two deformable registration algorithms might be that 3x3x3 voxel neighborhood is too small. The difference in registration quality is more apparent in the neighborhoods that contain edges between tissue types, and less apparent in the homogeneous neighborhoods. The smaller the neighborhood is the more likely it is to be homogeneous and therefore not to reflect the registration quality differences between the images. Thus, mutual information scores computed between such neighborhoods do not reflect the desired similarity between the images. As the neighborhood size increases the number of such homogeneous neighborhoods decreases and mutual information scores become more meaningful. Tables 7 and 8 show that it is indeed the case, that the number of voxels where the affine registration algorithm (FLIRT) has significantly higher mutual information scores than the two deformable registration algorithms drastically decreases as the size of the neighborhood increases. Our experiments suggest that at least 7x7x7 voxel neighborhoods should be used for voxel-wise comparison because 3x3x3 voxel neighborhood was not sensitive enough to reflect registration quality differences between FLIRT and the two fully deformable registration algorithms (Tables 7 and 8).

7 Summary

In summary, we propose a mutual information based methodology for comparing registration algorithms, and apply this method for quantitative performance evaluations on three registration algorithms. Our results show that both HAMMER and FEM-Demons algorithms perform better than FLIRT at global, voxel-wise and anatomical structure levels. Comparison between HAMMER and FEM-Demons yields that FEM-Demons has significantly higher mutual information scores on the global scale for MCI and Alzheimers patients (Table 1) while no significant difference is observed on controls. Voxel-wise comparison between the registration algorithms indicates that there are roughly twice as many voxels at which FEM-Demons significantly outperforms HAMMER than voxels where HAMMER outperforms FEM-Demons (Table 2). At the anatomical structure level, FEM-Demons and HAMMER produce significantly better results than FLIRT on almost all 116 anatomical structures. The difference in performance between FEM-Demons and HAMMER is the greatest for MCI patients, where FEM-Demons outperformed HAMMER on 31 anatomical structures, while HAMMER outperformed FEM-Demons on 15 (Table 3). HAMMER tends to do better on the anatomical structures with well-defined boundaries, like the ventricles and posterior cingulate, while FEM-Demons produces better results within homogeneous regions, like superioir temporal lobe and $cerebellum\_2b$. These results suggest that the choice between FEM-Demons and HAMMER should depend on the specific region of interest, because FEM-Demons outperforms HAMMER on some anatomical regions, while HAMMER outperforms FEM-Demons on others.

8 Appendix

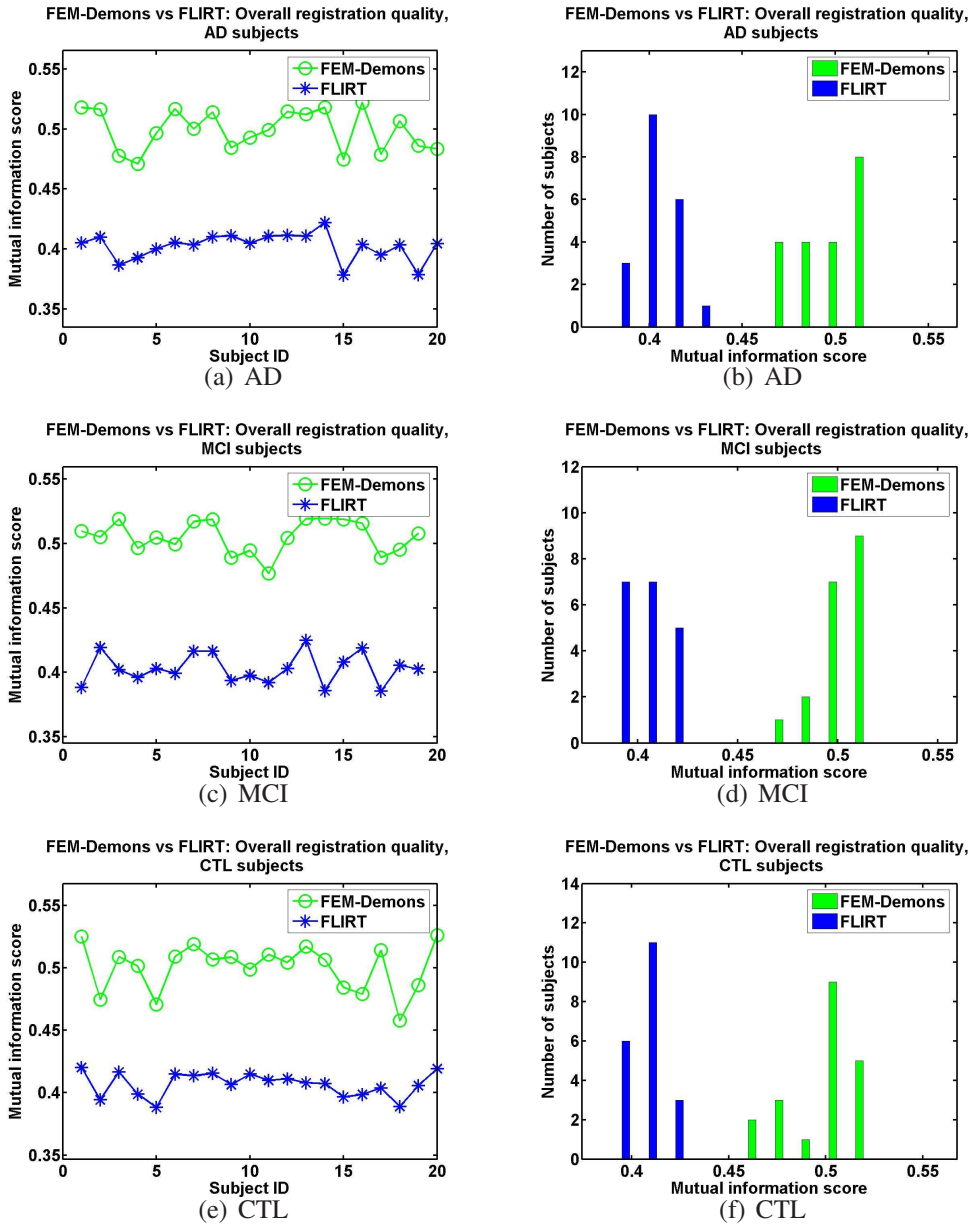


Figure 7: Global mutual information registration scores for FEM-Demons and FLIRT. a, b - Alzheimers patients; c, d - MCI patients; e, f - controls. Figures b, d, f show histograms of the MI score distribution between the two methods.

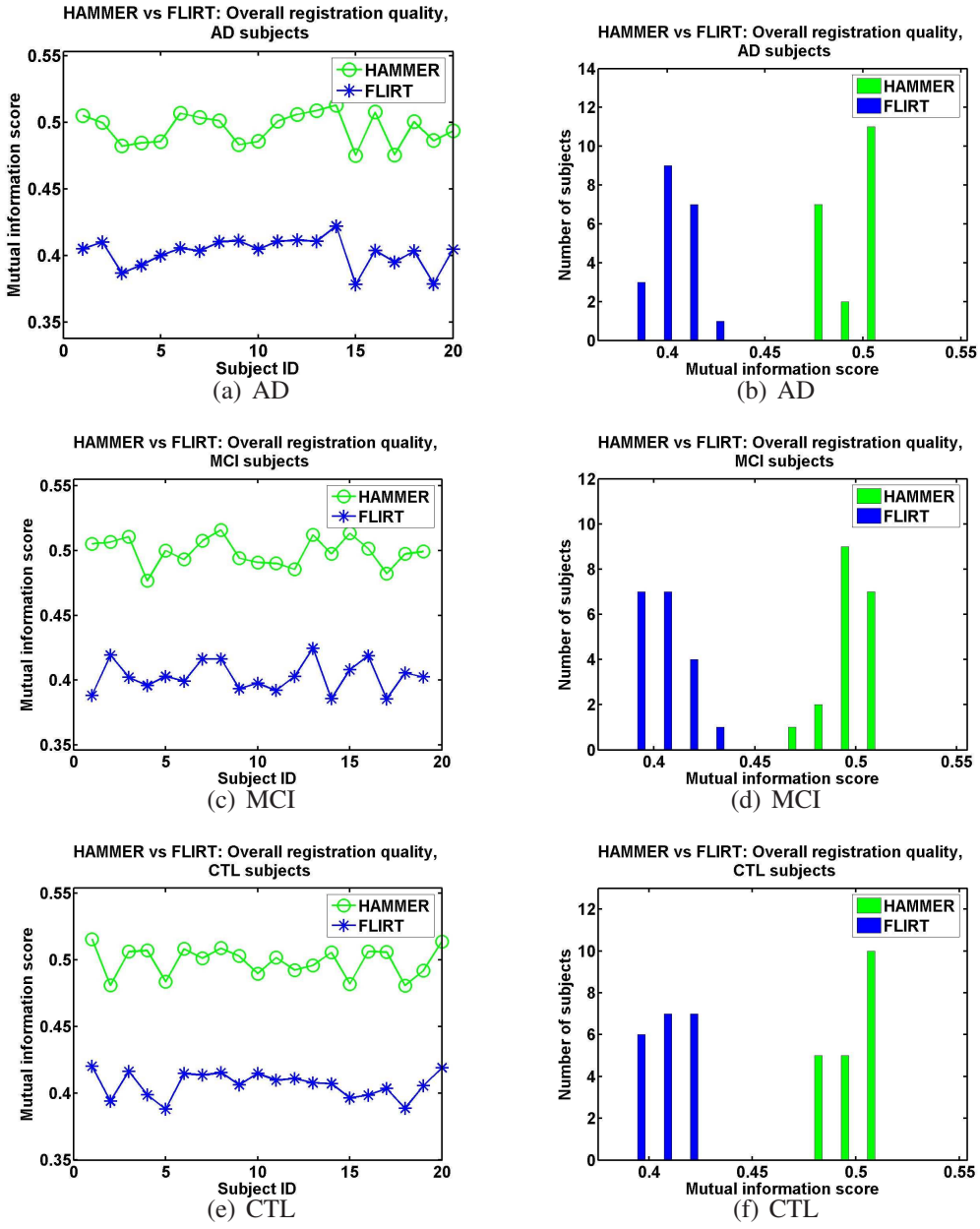


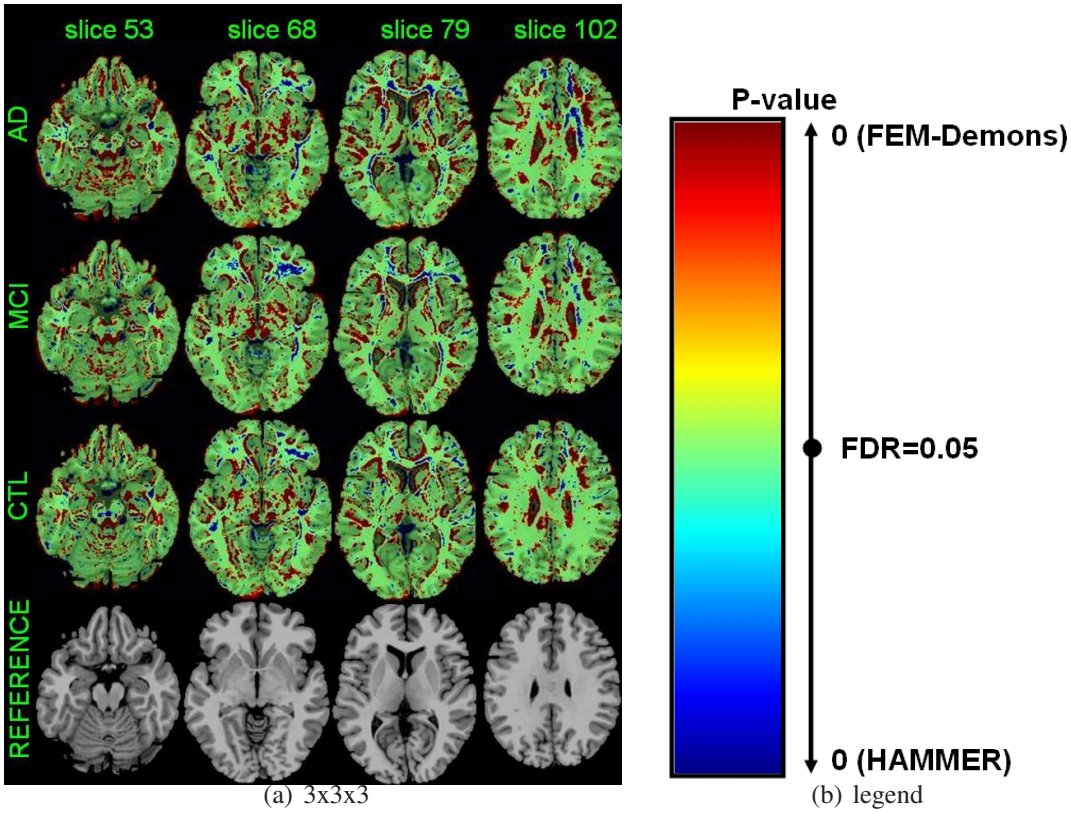
Figure 8: Global mutual information registration scores for HAMMER and FLIRT. a, b - Alzheimers patients; c, d - MCI patients e, f - controls. Figures b, d, f show histograms of the MI score distribution between the two methods.

Table 7: FEM-Demons vs FLIRT: percentage of voxels where mutual information scores for one method are significantly higher than for the other.

3x3x3			
Registration Method	Alzheimers	MCI	Controls
FEM-Demons			
FDR=0.01	12.02%	9.71%	10.38%
FDR=0.05	25.27%	21.83%	22.37%
FDR=0.10	40.22%	36.12%	36.22%
FLIRT			
FDR=0.01	2.04%	2.04%	2.11%
FDR=0.05	4.83%	5.09%	5.25%
FDR=0.10	8.99%	9.71%	10.03%
7x7x7			
Registration Method	Alzheimers	MCI	Controls
FEM-Demons			
FDR=0.01	42.61%	40.60%	40.60%
FDR=0.05	57.50%	55.43%	55.30%
FDR=0.10	68.76%	67.06%	66.86%
FLIRT			
FDR=0.01	1.79%	1.38%	1.32%
FDR=0.05	3.30%	2.73%	2.59%
FDR=0.10	5.25%	4.61%	4.42%
11x11x11			
Registration Method	Alzheimers	MCI	Controls
FEM-Demons			
FDR=0.01	78.00%	79.72%	78.34%
FDR=0.05	87.18%	88.58%	87.93%
FDR=0.10	91.86%	92.96%	92.87%
FLIRT			
FDR=0.01	0.54%	0.36%	0.33%
FDR=0.05	0.83%	0.61%	0.54%
FDR=0.10	1.21%	0.89%	0.81%

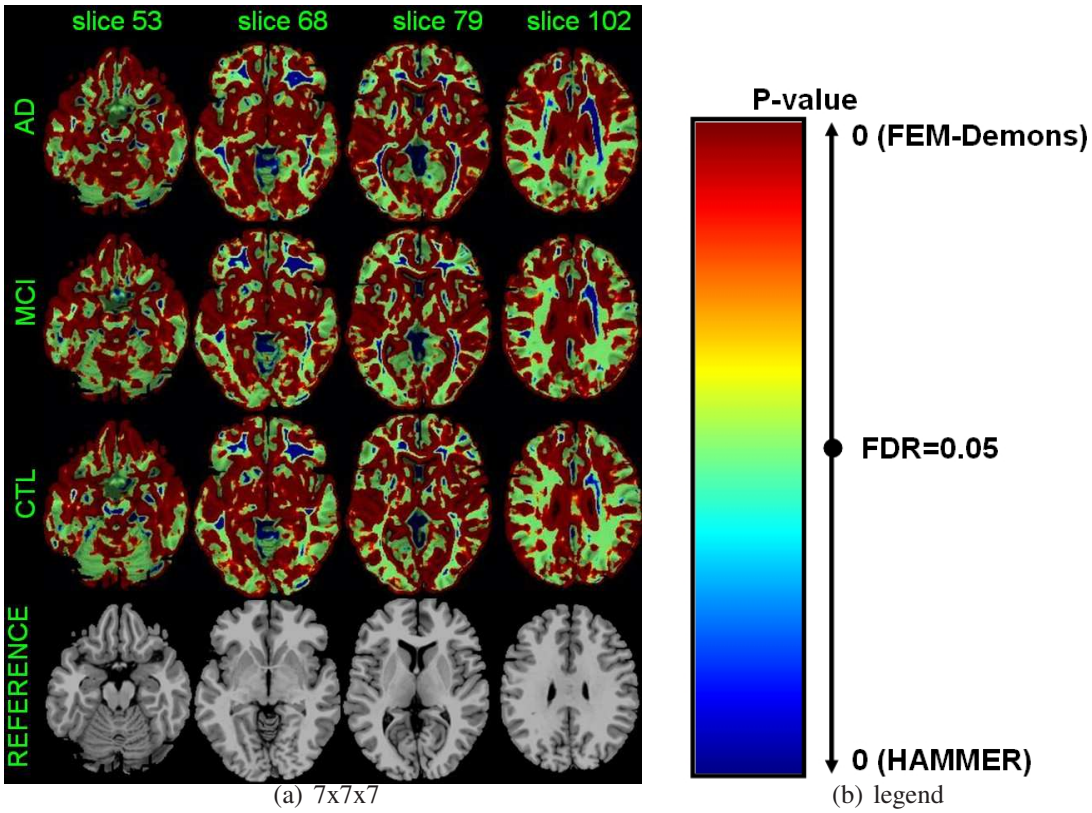
Table 8: HAMMER vs FLIRT: percentage of voxels where mutual information scores for one method are significantly higher than for the other.

3x3x3			
Registration Method	Alzheimers	MCI	Controls
HAMMER			
FDR=0.01	8.76%	7.56%	8.58%
FDR=0.05	18.29%	17.04%	18.48%
FDR=0.10	30.33%	29.49%	30.83%
FLIRT			
FDR=0.01	3.46%	2.21%	2.48%
FDR=0.05	8.02%	6.00%	6.30%
FDR=0.10	14.31%	11.88%	12.23%
7x7x7			
Registration Method	Alzheimers	MCI	Controls
HAMMER			
FDR=0.01	36.38%	35.16%	38.38%
FDR=0.05	50.08%	50.56%	52.75%
FDR=0.10	60.68%	62.32%	63.57%
FLIRT			
FDR=0.01	3.59%	1.96%	2.03%
FDR=0.05	6.11%	4.10%	4.16%
FDR=0.10	9.19%	6.93%	6.96%
11x11x11			
Registration Method	Alzheimers	MCI	Controls
HAMMER			
FDR=0.01	80.09%	82.81%	83.91%
FDR=0.05	88.07%	90.29%	90.35%
FDR=0.10	92.11%	93.83%	93.66%
FLIRT			
FDR=0.01	0.89%	0.46%	0.41%
FDR=0.05	1.33%	0.82%	0.76%
FDR=0.10	1.76%	1.20%	1.14%



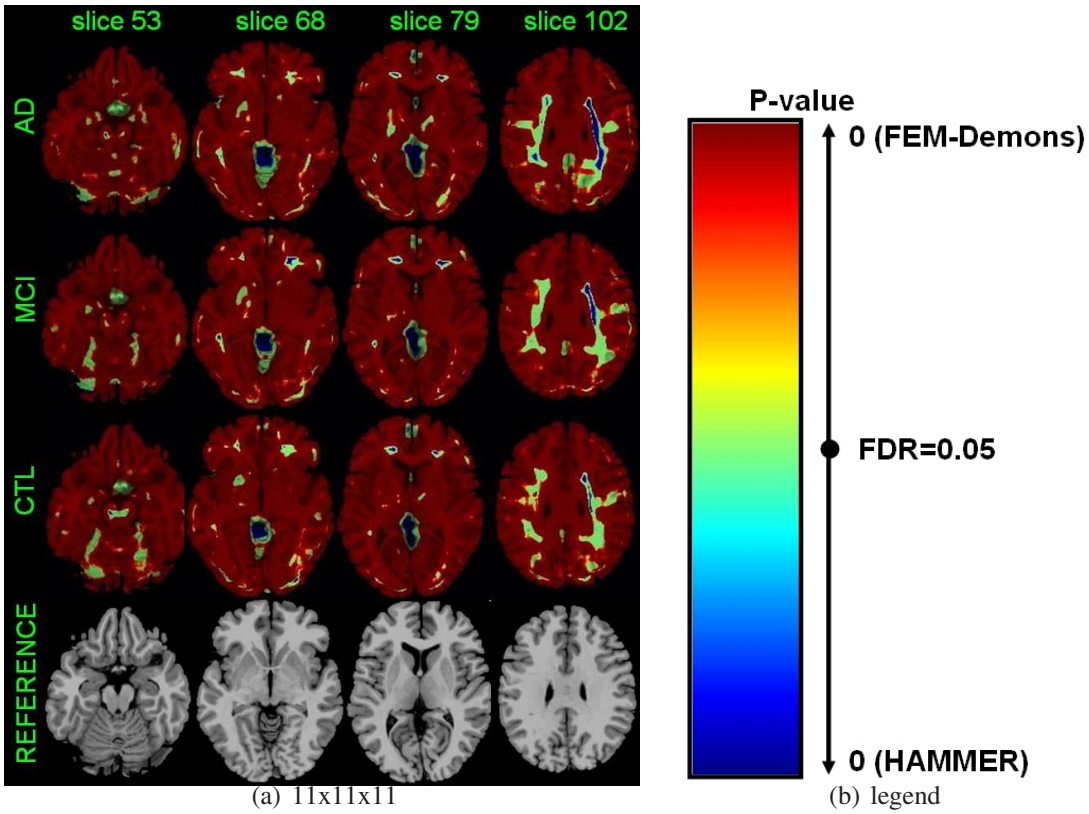
(a) 3x3x3 (b) legend

Figure 9: P-values from a paired T-test for every voxel. Red color indicates voxels where MI scores for FEM-Demons registration are significantly higher than that of FLIRT; blue color - voxels where FLIRT scores were better. Intensity of colors indicates statistical significance. (a) p-values for sample slices for 3x3x3 neighborhood; (b) shows the color map)



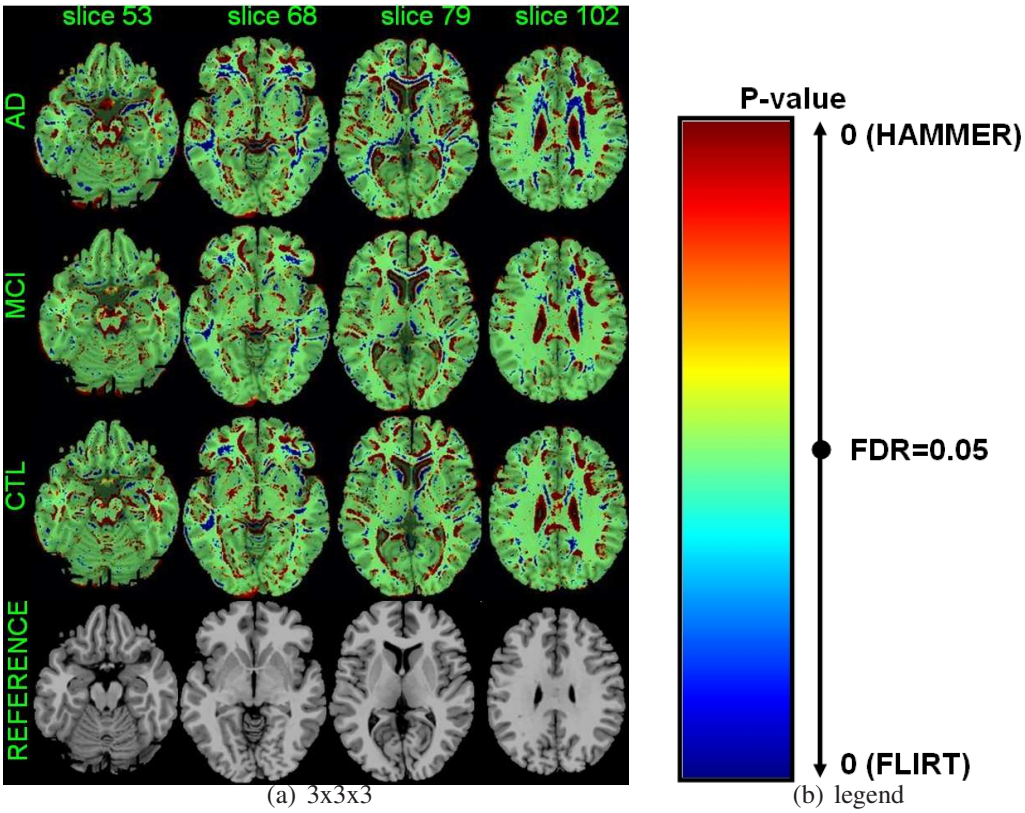
(a) 7x7x7 (b) legend

Figure 10: P-values from a paired T-test for every voxel. Red indicates voxels where MI scores for FEM-Demons registration are significantly higher than that of FLIRT; blue - voxels where FLIRT scores are higher. Intensity of colors indicates statistical significance. (a) p-values for sample slices for 7x7x7 neighborhood; (b) shows the color map)



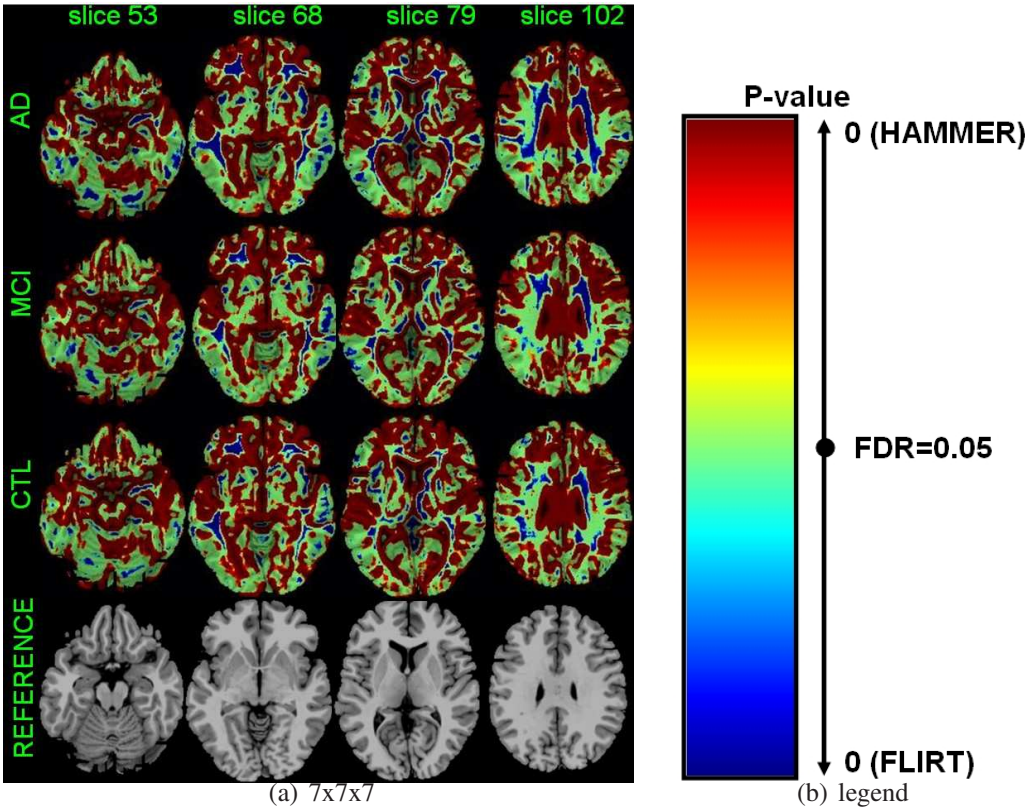
(a) 11x11x11 (b) legend

Figure 11: P-values from a paired T-test for every voxel. Red indicates voxels where MI scores for FEM-Demons registration are significantly higher than that of FLIRT; blue - voxels where FLIRT scores are higher. Intensity of colors indicates statistical significance. (a) p-values for sample slices for 11x11x11 neighborhood; (b) shows the color map)



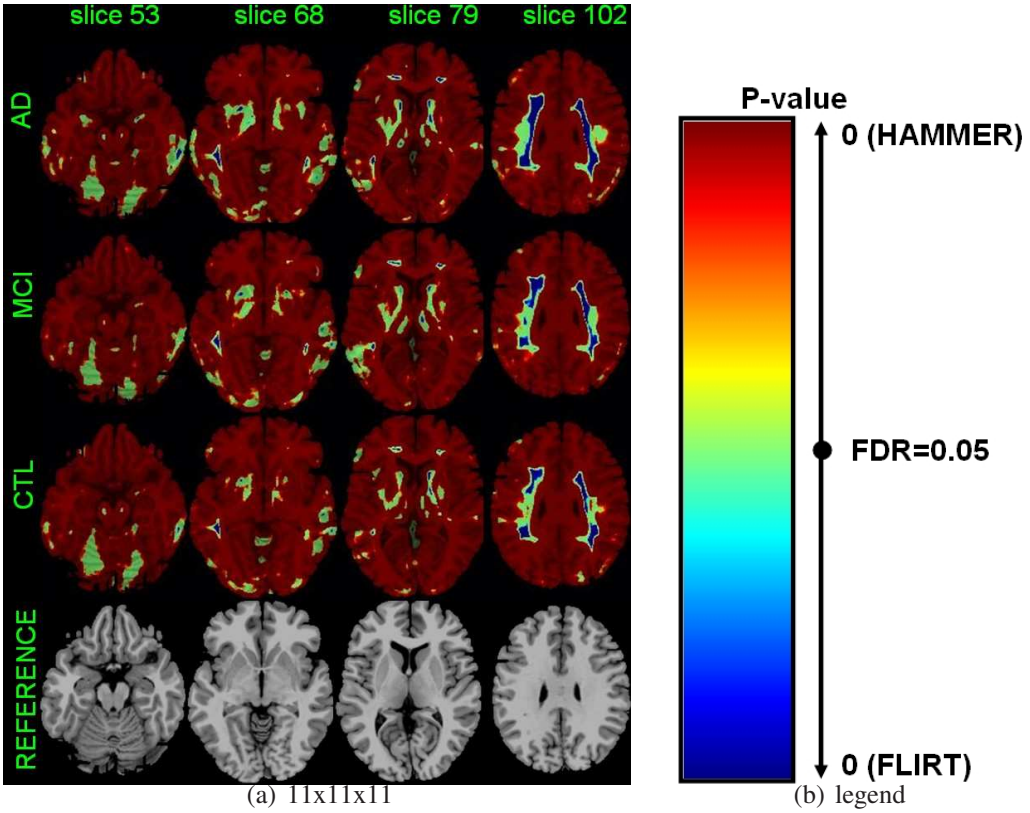
(a) 3x3x3 (b) legend

Figure 12: P-values from a paired T-test for every voxel. Red indicates voxels where MI scores for HAMMER registration are significantly higher than that of FLIRT; blue - voxels where FLIRT scores are better. Intensity of colors indicates statistical significance. (a) p-values for sample slices for 3x3x3 neighborhood; (b) shows the color map)



(a) 7x7x7 (b) legend

Figure 13: P-values from a paired T-test for every voxel. Red indicates voxels where MI scores for HAMMER registration are significantly higher than that of FLIRT; blue - voxels where FLIRT scores are higher. Intensity of colors indicates statistical significance. (a) p-values for sample slices for 7x7x7 neighborhood; (b) shows the color map)



(a) 11x11x11 (b) legend

Figure 14: P-values from a paired T-test for every voxel. Red indicates voxels where MI scores for HAMMER registration are significantly higher than that of FLIRT; color - voxels where FLIRT scores are higher. Intensity of colors indicates statistical significance. (a) p-values for sample slices for 11x11x11 neighborhood; (b) shows the color map)

References

[1] Benjamini, Y., Hochberg Y., *Controlling the false discovery rate: A practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society, Series B, Methodological, 57:289-300, 1995.

[2] O.T. Carmichael, H.A. Aizenstein, S.W. Davis, J.T. Becker, P.M. Thompson, C.C. Meltzer, and Y. Liu, *Atlas-based hippocampus segmentation in Alzheimers disease and mild cognitive impairment* , NeuroImage 27 979 990, 2005

[3] I.M. Chakravart, R.G. Laha, and J. Roy, *Handbook of Methods of Applied Statistics, Volume I.*, John Wiley, 1967, pp. 392-394.

[4] A. Collignon, *Multi-modality medical image registration by maximization of mutual information*, Ph.D. thesis, Catholic University of Leuven, Leuven, Belgium, 1998.

[5] C. Davatzikos, A. Genc, D. Xu, S. M. Resnick, *Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy*, NeuroImage 14, 1361 1369, 2001.

Table 9: FEM-Demons vs FLIRT: number of anatomical regions where mutual information scores for one method are significantly higher than for the other. Total of 116 regions were considered, as segmented on the AAL atlas.

Registration Method	Alzheimers	MCI	Controls
FEM-Demons			
FDR=0.01	105 (90.52%)	105 (90.52%)	107 (92.24%)
FDR=0.05	107 (92.24%)	110 (94.83%)	109 (93.97%)
FDR=0.10	108 (93.10%)	110 (94.83%)	109 (93.97%)
FLIRT			
FDR=0.01	1 (0.86%)	1 (0.86%)	0 (0.00%)
FDR=0.05	2 (1.72%)	1 (0.86%)	1 (0.86%)
FDR=0.10	3 (2.59%)	1 (0.86%)	1 (0.86%)

Table 10: HAMMER vs FLIRT: number of anatomical regions where mutual information scores for one method are significantly higher than for the other. Total of 116 regions were considered, as segmented on the AAL atlas.

Registration Method	Alzheimers	MCI	Controls
HAMMER			
FDR=0.01	113 (97.41%)	108 (93.10%)	114 (98.28%)
FDR=0.05	114 (98.28%)	113 (97.41%)	114 (98.28%)
FDR=0.10	114 (98.28%)	113 (97.41%)	114 (98.28%)
FLIRT			
FDR=0.01	0 (0.00%)	0 (0.00%)	0 (0.00%)
FDR=0.05	0 (0.00%)	0 (0.00%)	0 (0.00%)
FDR=0.10	0 (0.00%)	0 (0.00%)	0 (0.00%)

[6] J.M. Fitzpatrick, J.B. West, *A blinded evaluation and comparison of image registration methods* , Proc. Workshop on Empirical Evaluation Techniques in Computer Vision, University of California at Santa Barbara, IEEE Computer Society Press, Los Alimitos, CA, 12-27 (Jul 1998)

[7] J. C. Gee and R. K. Bajcsy, *Elastic Matching: Continuum Mechanical and Probabilistic Analysis*, Chapter in *Brain Warping*, ed. A. W. Toga, Academic Press, 1998.

[8] J. C. Gee and D. R. Haynor, *Numerical Methods for High Dimensional Warps*, Chapter in *Brain Warping*, ed. A. W. Toga, Academic Press, 1998.

[9] C.R. Genovese, N.A. Lazar and T.E. Nichols *Thresholding of statistical maps in functional neuroimaging using the false discovery rate*. NeuroImage 15: 870–878, 2002

[10] C.D. Good, R.I. Scahill, N.C. Fox, J. Ashburner, et al, *Automatic differentiation of anatomical patterns in the human brain: validation with studies of degenerative dementias*, NeuroImage 17 (1), 29 46, 2002.

[11] Y. Hochberg, A.C. Tamhane, *Multiple Comparison Procedures*. Wiley, 1987.

[12] C.J. Holmes, R. Hoge, L. Collins, R. Woods, A.W. Toga, A.C. Evans, *Enhancement of MR images using registration for signal averaging*, J Comput Assist Tomogr. 1998 Mar-Apr;22(2):324-33.

[13] L. Ibanez, W. Schroeder, L. Ng, J. Cates, *ITK Software Guide* Kitware, Inc., 2005.

[14] M. Jenkinson, S. Smith, *A global optimisation method for robust affine registration of brain images* Med Image Anal 2001;5(2):14356, http://www.fmrib.ox.ac.uk/fsl/flirt

[15] J. Klemencic, V. Valencic, N. Pecaric, *Deformable contour based algorithm for segmentation of the hippocampus from MRI*, Proceedings of the International Conference on Computer Analysis of Images and Patterns, 2001.

[16] Z. Lao, D. Shen, Z. Xue, B. Karacali, S. Resnick, and C. Davatzikos, *Morphological classification of brains via high-dimensional shape transformations and machine learning methods*, NeuroImage, Vol 21(1), pp 46-57, 2004.

[17] O.L. Lopez, J.T. Becker, W. Klunk, J. Saxton, R.L. Hamilton, D.I. Kaufer, R. Sweet, C.C. Meltzer, S. Wisniewski, M.I. Kamboh, S.T. DeKosky, *Research evaluation and diagnosis of probable Alzheimers disease over the last two decades: I*, Neurology 55, 1854 1862, 2000a

[18] F. Maes, A. Collignon, D. Vandermeulun, G. Marchal, and P. Suetens, *Multimodality image registration by maximization of mutual information*, IEEE Transactions on Medical Imaging, 16(2):187,198, 1997.

[19] D. Mattes, D.R. Haynor, H. Vesselle, T.K. Lewellen, and W. Eubank, *Nonrigid multimodality image registration*, Medical Imaging: Image Processing, M. Sonka and K. M. Hanson, Eds. 2001, vol. 4322 of Proc. SPIE, pp. 16091620, SPIE Press, Bellingham, WA.

[20] J. Mazziotta, A. Toga, A. Evans, P, Fox, J. Lancaster, K. Zilles, R. Woods, et al., *A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)*, Philos Trans R Soc Lond B Biol Sci. 2001 Aug 29;356(1412):1293-322.

[21] Y. Nichols, S. Hayasaka, *Controlling the Familywise Error Rate in Functional Neuroimaging: A Comparative Review*, Stat. Meth. in Med. Research, 12(5): 419-446, 2003.

[22] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipies in C*, Cambridge University Press, second edition, 1995

[23] R. Schestowitz, C. Twining, T. Cootes, V. Petrovic, C. Taylor, W. Crum, *Assessing the Accuracy of Non-Rigid Registration With and Without Ground Truth*, Proc. IEEE International Symposium on Biomedical Imaging, 2006

[24] C. E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal, vol. 27, pp. 379423/623656, 1948.

[25] D. Shen, C. Davatzikos, *HAMMER: Hierarchical Attribute Matching Mechanism for Elastic Registration*, IEEE Trans. on Medical Imaging, 21(11):1421-1439, Nov 2002.

[26] D.G. Shen , C. Davatzikos, *Very high resolution morphometry using mass-preserving deformations and HAMMER elastic registration.*, NeuroImage 18 (1), 28 41, 2003.

[27] D. Shen, S. Moffat, S.M. Resnick, C. Davatzikos. *Measuring size and shape of the hippocampus in MR images using a deformable shape model* NeuroImage 15 (2), 422 434 (February), 2002.

[28] S. Smith, *Fast robust automated brain extraction*, Hum. Brain Mapp. 17 (3), 143 155, 2002.

[29] J. P. Thirion, *Fast non-rigid matching of 3D medical image*, Technical report, Research Report RR-2547, Epidure Project, INRIA Sophia, May 1995.

[30] J. P. Thirion, *Image matching as a diffusion process: an analogy with maxwells demons*, Medical Image Analysis, 2(3):243260, 1998.

[31] P.M. Thompson, M.S. Mega, R.P. Woods, C. Zoumalan, et al., *Cortical change in Alzheimers disease detected with a disease-specific population based Brain Atlas*, Cereb. Cortex 11 (1), 1 16, 2001.

[32] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, et al.,*Automated anatomical labelling of activations in spm using a macroscopic anatomical parcellation of the MNI MRI single subject brain*, NeuroImage, 2002

[33] P. A. Viola, *Alignment by maximization of mutual information*, Ph.D. thesis, Massachusetts Institute of Technology, Boston, MA, USA, 1995.

[34] H.Wang, L.Dong, J.ODaniel, R.Mohan, A.S. Garden, K.K. Ang, D.A. Kuban, M. Bonnen, J.Y. Chang and R. Cheung, *Validation of an accelerated demons algorithm for deformable image registration in radiation therapy* , Phys. Med. Biol. 50 28872905, 2005.

[35] 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 Trans. on Medical Imaging, 20(1):45-57, 2001.

[36] P. Zhilkin, M.E. Alexander, *Affine registration: a comparison of several programs*, Magnetic Resonance Imaging 22(1):55-66, 2004.

[37] M. Wu, O. Carmichael, C. S. Carter, J. L. Figurski, P. Lopez-Garcia, H. J. Aizenstein, *Quantitative comparison of neuroimage registration by air, spm, and a fully deformable model*. In press, Human Brain Mapping, September 2005.

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Carnegie Mellon