A Framework for Detailed Objective Comparison of Non-rigid Registration Algorithms in Neuroimaging

William R. Crum[1], Daniel Rueckert[2], Mark Jenkinson[3], David Kennedy[4], and Stephen M. Smith[3]

[1] Computational Imaging Science Group, Division of Imaging Sciences, Thomas Guy House, Guy's Hospital, Kings College London, London SE1 9RT, UK, bill.crum@kcl.ac.uk
[2] Department of Computing, Imperial College, London SW7 2AZ, UK
[3] Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK
[4] Center for Morphometric Analysis, Department of Neurology, MGH, 13th Street, Charlestown MA 02129, USA

Abstract. Non-rigid image registration is widely used in the analysis of brain images to the extent it is provided as a standard tool in common packages such as SPM. However the performance of algorithms in specific applications remains hard to measure. In this paper a detailed comparison of the performance of an affine, B-Spline control-point and viscous fluid registration algorithm in inter-subject brain registration is presented. The comparison makes use of highly detailed expert manual labellings of a range of structures distributed in scale and in location in the brain. The overall performance is B-Spline, fluid, affine (best first) with all algorithms struggling to improve the match of smaller structures. We discuss caveats, evaluation strategies for registration and implications for future registration-based neuroimaging studies.

1 Introduction

Image registration is now widely used in medical image analysis but detailed comparison of the performance and suitability of algorithms for specific applications remains difficult, especially in the non-rigid case. In brain image analysis a common task is to register an MRI brain scan of one subject to another or to a template image in standard anatomical space. This is challenging due to population variability in cortical anatomy and the poorly specified nature of anatomical versus functional correspondence [1]. Many registration algorithms exist which can attempt this task but one persistent problem is how to conduct a detailed evaluation of such algorithms under realistic conditions.

In this paper we describe an application-centric framework for comparison and use it to evaluate the relative success of applying an affine [2], B-Spline [3] and viscous fluid registration (e.g. [4]) to the problem of inter-subject brain registration. This comparison framework makes use of highly detailed expert labelling of neuroanatomical structures on a set of test images [5], [6]. These labels allow two assessments to be made: (i) an assessment of the relative performance of the registration algorithms in aligning specific structures or tissue classes (ii) a generic assessment of

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3216, pp. 679–686, 2004.

which structures fail to be well registered by any algorithm. The latter case has important implications for the efficacy of large-scale neuroimaging studies where the identification of structural or functional differences relies on non-rigid registration. The evaluation is "application-centric" in that it tests the ability of each algorithm to bring a series of neuroanatomical structures into alignment, rather than testing the absolute correctness of the algorithms in recovering a known transformation between two data-sets. We put no conditions on how the registration algorithms were applied beyond suggesting that default or typical parameter settings should be used and that the registrations should be run once using these settings. We are assessing suitability for the task rather than performing a detailed technical comparison of algorithms.

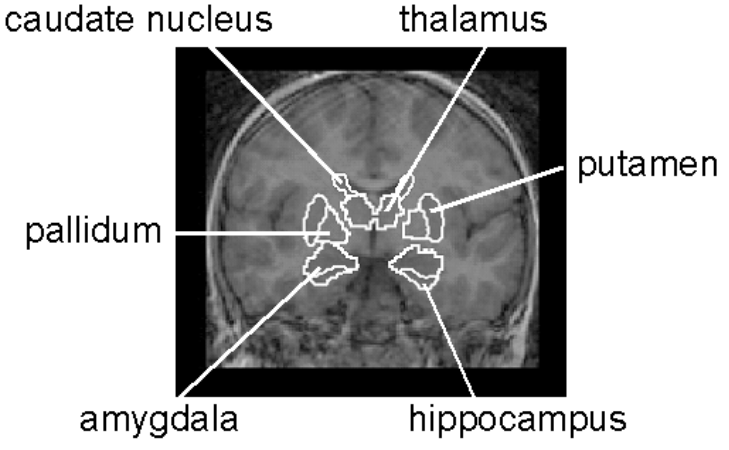

Fig. 1. A few of the sub-cortical structures available in the CMA brain data.

2 Method

2.1 Evaluation Data

We made use of eight labelled MR brain images obtained from the Centre of Morphometric Analysis at MGH (Boston). This centre has spent many years performing detailed reproducible manual labelling of MRI brain images [5], [6]. Each voxel has associated binary labels that identify it as a member of particular structures or tissue classes; there are 84 sub-cortical labels and 48 cortical labels. These brain images and a subset of the labels are available for use by the research community from the Internet Brain Segmentation Repository (http://www.cma.mgh.harvard.edu/ibsr/). For this work, the original labels have been grouped to produce a smaller hierarchy ranging from the entire brain and the primary lobes down to structures such as the hippocampus and the thalamus (see Table 1 for the full list of labels and Fig. 1 for some examples). The chosen groupings are arbitrary and can be defined to suit any specific application. In this paper we refer to the set of grouped labels as the test labels.

Table 1. The four groups of test labels used in the evaluation. The number of original anatomical labels used to create each test label is shown in brackets.

Major Structures	Major Lobes	Other Structures	Sub-Cortical
All (32)	Frontal (14)	Lat. Ventricle (2)	Thalamus (2)
Brain (25)	Occipital (8)	Cerebellum (4)	Caudate (2)
Cortex (6)	Parietal (7)	Brain Stem (3)	Putamen (2)
White Matter (2)	Temporal (16)	Sub-Cortex (10)	Pallidum (2)
CSF (7)			Hippocampus (2)
			Amygdala (2)

2.2 Evaluation Algorithms

Three non-rigid registration algorithms were evaluated. These were (a) an affine registration algorithm (FLIRT) available as part of the FSL image analysis toolkit (www.fmrib.ox.ac.uk/fsl) [2], [7] (b) a free-form deformation algorithm based on B-Splines [3], [8] and (c) a viscous fluid algorithm implemented following [4] and [9].

FLIRT: The affine registration method, FLIRT (FMRIB's Linear Registration Tool), was chosen as robust affine registration is widely used to register brain images and affine transformations are sufficient to successfully align many brain structures between individuals despite the relatively small number of degrees of freedom in the transformation model. It is designed to be highly robust to the initial alignment of the images by using a customised global optimisation method that runs over multiple scales (8, 4, 2 and 1mm); a large search space in rotation and scale is used at the 8mm resolution and many smaller perturbations on the best three candidate solutions used in the 4mm resolution. In addition, the cost functions are regularised such that they de-weight contributions near the edge of the overlapping field of view in order to produce a smoothly changing cost function. All of these factors combine to reduce the chance of the registration becoming "trapped" in a local minimum of the cost function. The correlation ratio is used to drive the registration.

B-SPLINE: Non-rigid registration based on free-form deformations and B-Splines is widely used in many registration applications including those involving inter-subject brain registration. The basic idea of FFDs is to deform an object by manipulating an underlying mesh of control points. The optimal control point locations are found by minimising a cost function which encompasses two competing goals: the first term represents the cost associated with the voxel-based similarity measure, in this case normalised mutual information, while the second term corresponds to a regularisation term which constrains the transformation to be smooth. The control-point spacing was set at 2.5mm for this study.

FLUID: The fluid methods have been used successfully for intra-subject brain registration [10] [11] and for registering structures from one brain to another [4]. They use a mathematical model of a compressible viscous fluid to model the transformation between images. They can accommodate large deformations but can be less robust than other methods without good initialisation. Therefore the fluid algorithm was initialised from a locally derived affine registration of each subject into a standard ana-

tomical space. Additionally the registration was terminated after 5 regridding steps to reduce the influence of numerical error. The intensity cross correlation was used to drive the registration as in [9].

2.3 Evaluation Measures

We observe that for most applications of non-rigid registration in neuroimaging it is correspondence of brain structures on a variety of scales that is important as the absolute correctness of the transformation model cannot usually be determined; a corollary is that the amount of information contained in an MR brain image is insufficient to assess the point-accuracy of the registration but anatomical structures can be tested for correspondence post registration. Therefore the test labels that we have generated from the CMA data represent a natural means for evaluation of registration algorithms that is closely tied to their uses for neuroimaging research. Given pairs of test-labels, S and T, on registered brains a method for evaluating their overlap is required. There is considerable literature in this area, much of it applied to the assessment of segmentation algorithms (e.g. [12]). In this work we use P, the ratio of the number of overlapping voxels to the total number of voxels in the labelled structures.

$$P = \frac{N(S \cap T)}{N(S \cup T)} \tag{1}$$

This measure has been widely used but has the known disadvantage that errors in labelling small structures are magnified compared with larger structures.

2.4 Evaluation Framework

Each of the eight test subjects was registered to the other seven subjects using each of the three algorithms giving 56 inter-subject registrations for each algorithm. The registrations were run by the researchers most familiar with their operation; these researchers made any necessary parameter choices independently. After registration, each of the 19 test labels on each subject was transformed into the space of all the other subjects using the transformations determined by each registration algorithm. The binary labels were transformed using trilinear interpolation and then thresholded at 50% to produce transformed binary labels. The test-label overlap, P, was computed in all cases and this data was analysed to produce a mean and standard deviation fractional overlap for each test-label for each registration algorithm. The overlaps were also computed for all pairs of unregistered scans for comparison.

3 Results

The FLIRT registrations took approximately 5 minutes each to run on a contemporary desk-top Linux PC. Both the B-Spline and fluid registrations were run in a distributed fashion on a Linux condor cluster (one CPU per registration) and typically took be-

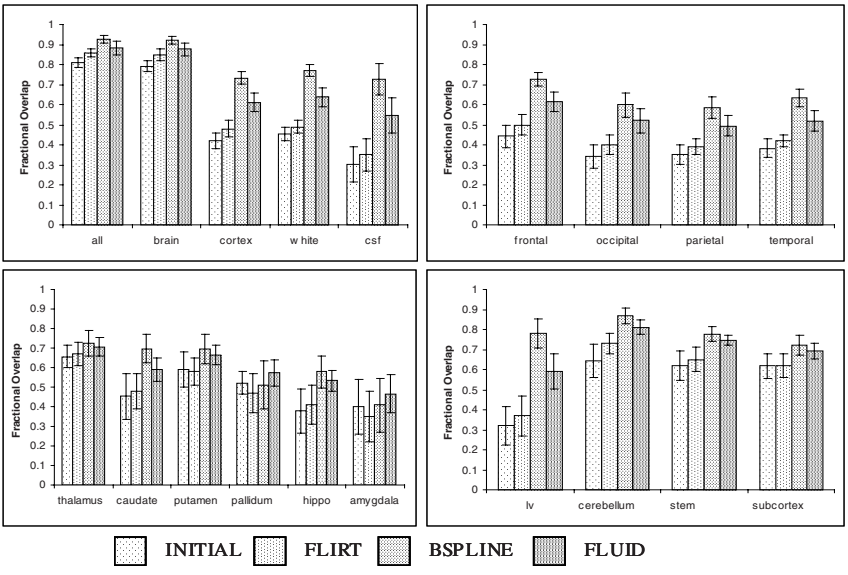

Fig. 2. The mean and standard deviation of the overlap measure for each structure registered by each registration algorithm. INITIAL refers to the original images.

tween 2 and 10 hours per registration. The results are summarised in Fig. 2. For each of the test labels, the mean fractional overlap is shown for each registration algorithm and the error bars show ±1 s.d. Some general observations can be made. In all cases all registration algorithms improved the mean label overlap except for the case of FLIRT applied to the pallidum and amygdala. We might expect that the potential accuracy of matching increases with the degrees of freedom available to the algorithms but that the potential for mis-registration also increases. In fact what we found is that for this study, the B-Spline method consistently performed well and was only outperformed by the fluid method for two structures (the pallidum and amygdala) where all methods struggled to improve the overlap. Conversely the B-Spline method proved far superior to all other methods for registering the larger tissue compartments (cortex, white, csf and lateral ventricle). There are some caveats associated with these results. The most obvious is that the sample-size of subject data is relatively small but another important point is that as part of the labelling process, the subject data was realigned and had already been interpolated prior to our analysis. This might have the largest impact on the alignment of small structures.

4 Discussion

We have assessed the ability of three registration algorithms to align a variety of brain structures between eight subjects and found overlaps ranging from ~0.3 (lateral ventricle) to ~0.9 (brain). The results enable us to distinguish the performance of the algorithms over different parts of the brain. Notable is that the B-Spline algorithm

matches CSF and lateral ventricle particularly well, no algorithms match the major lobes better than ~0.7 and that there is less difference between the algorithms ability to match the sub-cortical structures than a consideration of the degrees of freedom of the transformation model used in each case might suggest. We would normally expect the largest increase in overlap to be achieved by the affine registration compared with no registration however due to the realignment applied to the images prior to manual segmentation the observed increases are relatively small. The fact that each algorithm used a different image similarity measure must impact on the results but reflects the current lack of consensus and deep understanding of the operation of such measures.

The most relevant recent work on registration evaluation is Hellier *et al* [13] where six registration algorithms were compared using a variety of measures including tissue overlap, correlation of differential characteristics and sulcal shape characteristics. They found that algorithms with higher degrees of freedom did not perform proportionately better at matching cortical sulci and that inter-subject cortical variability remains a severe challenge for voxel-based non-rigid algorithms. These findings are consistent with our experience. Their choice of a single reference subject could be a source of bias that we avoid by registering all permutations of the test subjects. Previously, Grachev *et al* [14] suggested using 128 carefully defined and manually placed landmarks per hemisphere to evaluate inter-subject registration accuracy. This approach, like ours, requires significant operator expertise to identify features. Landmarks enable a millimetre error to be computed but labels allow assessment that is more easily related to correspondence of the underlying neuroanatomy. Label-based evaluation approaches can also estimate the degree to which non-rigid registration can align anatomically and functionally important areas of the brain. This might prove important for establishing error bounds in studies using registration to compare groups of individuals. Another application is in serial scanning of individuals where labels defined on the first scan in a sequence may be propagated to subsequent scans using non-rigid registration.

Simple voxel-intensity driven algorithms will not on their own resolve the outstanding problem of cortical variability between subjects but are likely to remain useful for matching the locale of similar cortical structures. This degree of matching may be perfectly adequate for many applications especially where explicit allowance is made for uncertainty in structural matching e.g. in Voxel Based Morphometry [15]. For more specialised applications hybrid approaches may be required; for example cortical matching may be improved - or at least better controlled – by exploiting work done by Maudgil *et al* [16] where 24 points which were determined to be homologous between subjects were identified on the cortical surface. Such points can be incorporated into voxel-based registration algorithms [17].

There remain some questions as to the best way to define label overlaps. The measure we use in this paper is well known but does not account explicitly for inconsistency in the labelling process nor does it indicate the nature of the error in non-perfect overlaps. Crum *et al* [1] suggest the use of a tolerance parameter with overlap measures so that, for instance with the tolerance set to 1 voxel, the boundaries of two labels are regarded as completely overlapping if they are at most one voxel away from each other. A deeper consideration of overlap measures and the incorporation of labelling error is an urgent priority.

We plan to extend this study by including more registration algorithms and more labelled subjects in the evaluation. Future work will focus on a more detailed technical comparison of algorithms but this initial work has provided a benchmark for fu-

ture performance in two ways. First, other registration techniques can be easily tested within the same framework for an operational comparison. Second, we can recognise that these algorithms are subject to many parameter choices that we have ignored in this study. We can optimise the parameter choice for each algorithm with respect to this well-defined task; this optimisation process may ultimately lead to new methods tuned to register structures of particular scale and intensity characteristics.

Acknowledgments. The authors are grateful for the intellectual and financial support of the Medical Images and Signals IRC (EPSRC and MRC GR/N14248/01). This work was made possible by the research environment created by this consortium. We also acknowledge valuable contributions from Professor David Hawkes and Dr Derek Hill.

References

1. Crum, W.R., Griffin, L.D., Hill, D.L.G. and Hawkes, D.J. : Zen and the Art of Medical Image Registration : Correspondence Homology and Quality. NeuroImage 20 (2003) 1425-1437
2. Jenkinson, M. and Smith, S.M. : A Global Optimisation Method for Robust Affine Registration of Brain Images. Medical Image Analysis 5(2) (2001) 143-156
3. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O. and Hawkes, D.J. : Non-Rigid Registration Using Free-Form Deformations: Application to Breast MR images. IEEE Transactions on Medical Imaging 18(8) (1999) 712-721
4. Christensen, G.E., Joshi S.C. and Miller, M.I. : Volumetric Transformation of Brain Anatomy. IEEE Transactions on Medical Imaging 16(6) (1997) 864-877
5. Kennedy, D.N., Fillipek, P.A. and Caviness, V.S. : Anatomic Segmentation and Volumetric Calculations in Nuclear Magnetic Resonance Imaging. IEEE Transactions on Medical Imaging 8 (1989) 1-7
6. Caviness, V.S., Meyer, J., Makris, N. and Kennedy, D.N. : MRI-based Topographic Parcellation of the Human Neocortex: an Anatomically Specified Method With Estimate of Reliability. Journal of Cognitive Neuroscience 8 (1996) 566-587
7. Jenkinson, M., Bannister, P., Brady, J.M. and Smith, S.M. : Improved Optimisation for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage 17(2) (2002) 825-841
8. Rueckert, D., Frangi, A.F. and Schnabel, J.A. : Automatic Construction of 3D Statistical Deformation Models of the Brain using Non-Rigid Registration. IEEE Transactions on Medical Imaging 22(8) (2003) 1014-1025
9. Freeborough, P.A. and Fox, N.C. : Modeling Brain Deformations in Alzheimer Disease By Fluid Registration of Serial 3D MR Images. Journal of Computer Assisted Tomography 22(5) (1998) 838-843
10. Crum, W.R., Scahill, R.I., Fox, N.C. : Automated Hippocampal Segmentation by Regional Fluid Registration of Serial MRI : Validation and Application in Alzheimer's Disease. NeuroImage 13(5) (2001) 847-855
11. Fox, N.C., Crum, W.R., Scahill, R.I., Stevens, J.M., Janssen, J.C. and Rossor, M.N. : Imaging of Onset and Progression of Alzheimer's Disease with Voxel-Compression Mapping of Serial MRI. The Lancet 358 (2001) 201-205
12. Gerig, G., Jomier, M. and Chakos, M. : Valmet: A new validation tool for assessing and improving 3D object segmentation. In : Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science, Vol. 2208, Springer-Verlag, Berlin Heidelberg New York (2001) 516-528

13. Hellier, P., Barillot, I., Corouge, B., Gibaud, G., Le Goualher, G., Collins, D.L., Evans, A., Malandain, G., Ayache, N., Christensen, G.E. and Johnson H.J. : Retrospective Evaluation of Intersubject Brain Registration. IEEE Transactions on Medical Imaging, 22(9) (2003) 1120-1130
14. Grachev, I.D., Berdichevsky D., Rauch, S.L., Heckers, S., Kennedy, D.N., Caviness V.S. amd Alpert, N.M. : A Method For Assessing the Accuracy of Intersubject Registration of the Human Brain Using Anatomic Landmarks, NeuroImage 9 (1999) 250-268
15. Ashburner, J., and Friston, K.J. : Voxel-Based Morphometry – the Methods. NeuroImage 11 (2000) 805-821
16. Maudgil, D.D., Free, S.L., Sisodiya, S.M., Lemieux, L., Woermann, F.G., Fish, D.R., and Shorvon, S.D. : Identifying Homologous Anatomical Landmarks On Reconstructed Magnetic Resonance Images of the Human Cerebral Cortical Surface. Journal of Anatomy 193 (1998) 559-571
17. Hartkens, T., Hill, D.L.G., Castellano-Smith, A.D., Hawkes, D.J., Maurer, C.R., Martin, A.J., Hall, W.A., Liu, H. and Truwit, C.L. : Using Points and Surfaces to Improve Voxel-Based Non-Rigid Registration. in : Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science, Vol. 2489, Springer-Verlag, Berlin Heidelberg New York (2002) 565-572