Multiclassifier Fusion in Human Brain MR Segmentation: Modelling Convergence

Rolf A. Heckemann[1], Joseph V. Hajnal[1], Paul Aljabar[2], Daniel Rueckert[2], and Alexander Hammers[3]

[1] Imaging Sciences Department, MRC Clinical Sciences Centre, Imperial College at Hammersmith Hospital Campus, London, UK
jo.hajnal@imperial.ac.uk
[2] Department of Computing, Imperial College London, UK
[3] Division of Neuroscience and Mental Health, MRC Clinical Sciences Centre, Imperial College at Hammersmith Hospital Campus, London, UK*

Abstract. Segmentations of MR images of the human brain can be generated by propagating an existing atlas label volume to the target image. By fusing multiple propagated label volumes, the segmentation can be improved. We developed a model that predicts the improvement of labelling accuracy and precision based on the number of segmentations used as input. Using a cross-validation study on brain image data as well as numerical simulations, we verified the model. Fit parameters of this model are potential indicators of the quality of a given label propagation method or the consistency of the input segmentations used.

1 Introduction

Established methods for segmenting magnetic resonance (MR) images of the human brain rely either on automatic tissue classification or on manual delineation of anatomical regions. Automating anatomical segmentation would enable exciting new image analysis applications in diagnostic radiology and imaging research, such as brain volumetry on cohorts.

Given an unsegmented target image and an atlas (a reference MR image with a corresponding set of manually generated labels), estimating spatial anatomical correspondence between the image pair makes it possible to adapt the segmentation of the atlas to the target image [1,2,3]. This process is sometimes referred to as label propagation. Errors in the reference (atlas) segmentation and the anatomical correspondence estimate affect the accuracy of the propagated label set. The precision of the method (i.e., agreement between propagated labels from multiple atlases onto a target) depends on the consistency with which the different atlases have been segmented.

* Rolf Heckemann was supported by a research fellowship from the Dunhill Medical Trust. The authors would like to thank the National Society for Epilepsy, Chalfont St Peter, Buckinghamshire, UK, for making available the original 30 MRI datasets used in the creation of the atlases. This work was funded by the UK EPSRC as part of the IXI Project.

R. Larsen, M. Nielsen, and J. Sporring (Eds.): MICCAI 2006, LNCS 4191, pp. 815–822, 2006.

By combining multiple segmentations of a single target using vote rule decision fusion [4], random errors in the atlas segmentation and registration tend to cancel, resulting in an improved target segmentation. This has been demonstrated for confocal microscopy images of bee brains [5] as well as for MR images of human brains (albeit with large atlas regions) [1]. Decision fusion has also been applied successfully to a method of parcellation of the human cortex [6]. The method can also be refined by differential weighting of the input classifiers based on expectation-maximization algorithms [7,8].

The aim of this work was to investigate the relationship between the number of equally weighted input classifiers and the segmentation improvement achieved. We hypothesized that the rate of convergence with increasing atlas numbers can be modelled using a characteristic equation. By fitting model parameters based on limited input data, the maximum achievable segmentation accuracy as well as the quality of the registration can be determined. We carried out a leave-one-out cross-validation study on 30 expertly segmented human brain MR image volumes. The effect of the input data upon the fitted model was explored using simulations.

2 Method

Data. Three-dimensional T1-weighted MR volumes were available from 30 normal volunteers, age range 20–54 years, median age 30.5 years, 15 male, 15 female, 25 strongly right-handed, 5 non-right-handed. Each data set was accompanied by a set of labels in the form of an annotated image volume, where every voxel was coded as one of 67 anatomical structures (or background, code 0). These labels had been prepared using a protocol for manually outlining anatomical structures on two-dimensional sections from the image volume. The protocol used was an augmented version of a published prototype [9].

Image Registration and Label Propagation. Every subject was paired with every other subject for image registration. Intracranial structures were extracted from the target MR images using "BET" [10]. All image pairs were aligned using 3D voxel-based registration, maximizing normalized mutual information [11] in three steps. Rigid and affine registration corrected for global differences. In the third, nonrigid step, alignment of details in the image pair was achieved by manipulating a free-form deformation represented by displacements on a grid of control points blended using cubic B-splines [12]. The spacing of control points defines the local flexibility of the nonrigid registration. It was carried out in a multi-resolution fashion using successive control point spacings of 20 mm, 10 mm, 5 mm and 2.5 mm.

The final transformation was applied to the atlas labels using nearest-neighbor interpolation, generating 29 propagated label volumes for each target individual. In addition, we generated propagated label volumes based on less detailed registration, using the intermediate (rigid, affine and nonrigid with a control point spacing of 20 mm) transformation output to propagate the label volumes.

Segmentation Comparison. As a measure of agreement between two segmentations of a target, we used the similarity index (SI). For a pair of labels, it

is defined as the ratio between the volume of the intersection (overlap) and the mean volume of a pair of labels in the same coordinate space [13]. To compare full label sets, we calculated the mean SI over all structures.

$$SI_m = \frac{1}{67} \sum_{k=1}^{67} \frac{2n(L_a^k \cap L_b^k)}{n(L_a^k) + n(L_b^k)} \tag{1}$$

$L_{a,b}^k$: Labels compared; k: structure code; n: Number of voxels. The measure ranges from 0 (for label sets that have no overlapping regions) to 1 (for label sets that are identical).

Decision Fusion Model. Label volumes represent classifiers that assign a structure label to every voxel in the corresponding MR image volume. To combine the information from multiple individual propagated label volumes into a consensus segmentation, the classifiers were fused on a per-voxel basis using vote rule decision fusion as described by Kittler et al. [4]. For each of the 30 target brain images, we created consensus (fused) segmentations from subsets of propagated label volumes of varied sizes ($n = \{3, 5, 7, .., 29\}$). Where possible (for $n \leq 13$), we used multiple non-overlapping subsets. Individually propagated segmentations and fused label volumes were then compared with the manual label volume to determine the behaviour of SI_m as a function of n. Similarly, fused label volumes from independent sets of classifiers were compared with each other.

As the number of input label volumes increases, the SI values are expected to increase from an initially low level resulting from the combined effects of both systematic and random errors, towards an asymptotic value that reflects only the systematic errors. Assuming a Gaussian distribution for SI, we expect it to evolve with the number (n) of classifiers fused according to the secular equation:

$$SI_m(n) = 1 - a - \frac{b}{\sqrt{n}} \tag{2}$$

where a and b are parameters to be determined. Parameter a determines the asymptotic upper limit and reflects systematic differences between the labels being compared, whereas b is related to the random variability of the propagated labels. A small value of a implies consistent labelling.

Numerical Simulations. Numerical simulations were performed using a two-dimensional model as follows: A filled circle of radius 10 mm in an image matrix with 1 mm × 1 mm pixels was defined to represent a ground truth label I_{orig}. A random free-form deformation was applied by overlaying a grid of control points with 8 mm spacing and displacing these control points. The displacements were drawn from a Gaussian distribution with a mean of zero and a standard deviation of σ_{sys}. The resulting label I_{sys} represents an approximation to the ideal circular label containing systematic error. The model label, I_{sys}, was then deformed multiple times with grid-point displacements drawn from a different Gaussian distribution with mean zero and standard deviation σ_{rand}, producing

an ensemble of shapes that contained random errors, representing propagated labels. In step 3, independent ensembles were fused to create labels ($I^n{}_{fused}$) that represented estimates of the original, circular label, subject to systematic and random error. The agreement (as measured by SI) of individual and fused labels with the original gold standard label and with other fused labels was then investigated to explore the behaviour of SI as a function of the number of classifiers fused.

3 Results

MR Data. The accuracy of individual propagated label sets as measured by their mean SI_m with the manually generated target label set was 0.754 (range 0.692–0.799, $SD = 0.016$, $n = 870$). Compared to this baseline result, decision fusion consistently improved the level of agreement with the manual sets. The maximum accuracy was achieved using 29 classifiers: 0.836 (range 0.820–0.853, $SD = 0.009$, $n = 30$). The relationship between the number of input classifiers and the agreement level was described by Equation 2, with 95 % confidence intervals not exceeding the uncertainty of the individual data points. Fit parameters were $a = 0.144$ and $b = 0.10$ (see Fig. 1, fused-manual).

The precision of segmentation-fusion as measured by mean SI_m between independently generated fused label sets was also dependent on the number of input classifiers for each fused set. For three classifiers, SI_m was 0.812 ($SD = 0.005$, $n = 270$). Adding classifiers resulted in progressive improvement of this precision measure, up to 0.908 ($SD = 0.003$, $n = 30$) for subsets containing the maximum possible odd number of 13 independent classifiers (see Fig. 1, fused-fused). Again, the model described the curve appropriately, albeit with a wider confidence margin. Parameter estimates were $a = 0.016$ and $b = 0.29$.

Model parameters were dependent on the level of detail considered by the image registration algorithm underlying the label propagation process. The model fit for coarse registrations resulted in higher values of the a parameter, indicating that the maximum achievable agreement level for infinite classifier numbers is lower. The b parameter was also higher, indicating that the approach to convergence as classifier numbers increase is slower when coarser registration is used (Fig. 2).

Simulated Label Data. The SI between the fused and the gold standard label was well approximated by the model for nearly all σ_{sys} and σ_{rand}. The model fit was worse for SI calculated between two fused labels, reproducing the finding in the experimental data. Fig. 3 shows a plot of SI versus n, where the random parameters were $\sigma_{sys} = 4$ and $\sigma_{rand} = 5$.

The variation of the a and b parameters with σ_{rand} is shown in Fig. 4. Parameter b was found to be linear over the range of values studied. Parameter a was negative for low values of σ_{rand}, indicating a failure of the Gaussian assumption as the SI values approached the limiting value of 1. For multiple simulated labels the distribution of SI values passed a Kolmogorov-Smirnov normality test ($\alpha = 0.01$, $n = 100$) once the level of systematic error (σ_{sys}) exceeded 1mm.

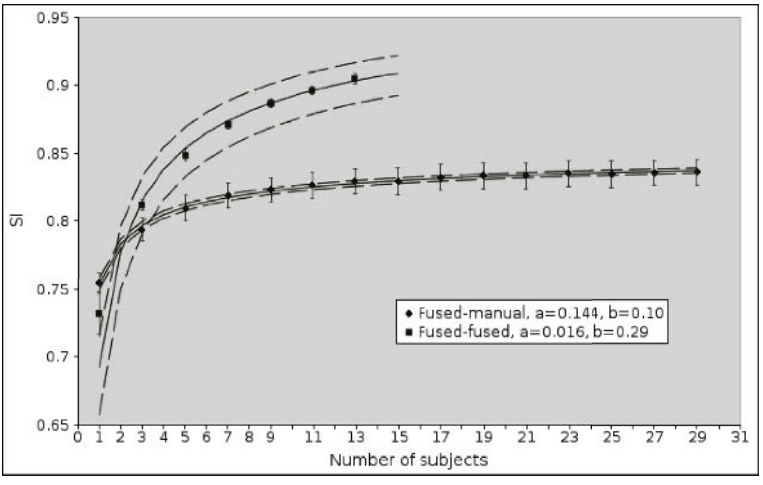

Fig. 1. *SI* vs. number of subjects in subset. Parameters a and b are shown as determined by model fitting. Error bars indicate standard deviation. Dashed lines indicate 95 % confidence intervals of the model fit. Agreement between fused labels is high with a poorer model fit.

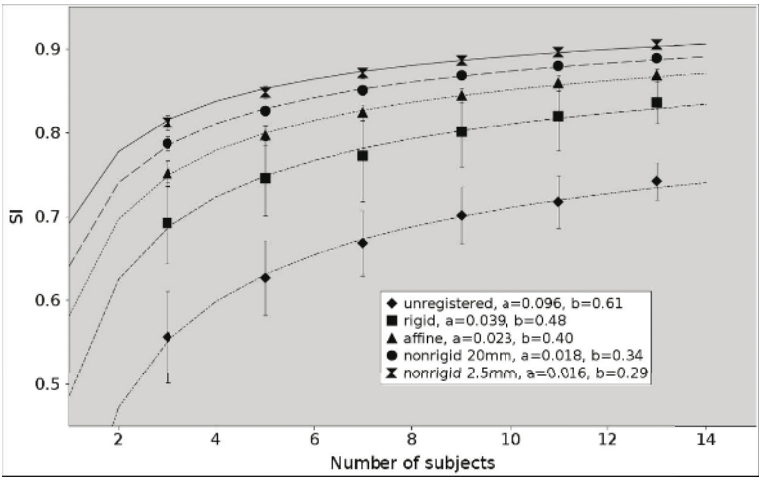

Fig. 2. Fused-to-fused comparisons: *SI* vs. number of subjects per subset (n) for different registration strategies. The a and b parameters are shown as determined by nonlinear model fitting.

4 Discussion and Conclusions

When multiple propagated segmentations of a brain image are regarded as classifiers and combined using a suitable decision fusion algorithm, the resulting fused segmentation can be more accurate than any of the constituent segmentations,

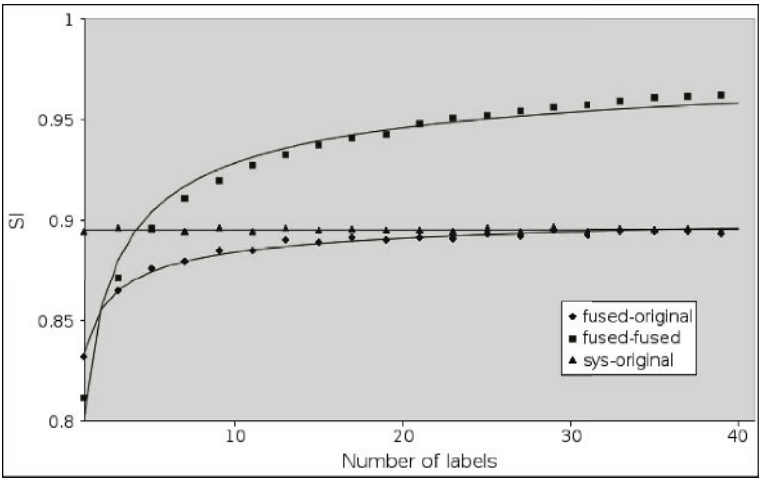

Fig. 3. Plots of SI versus n for various comparisons of simulated labels. The comparison between I_{fused} and I_{orig} is well fitted by the model. Overlaps between I_{sys} and I_{orig} are shown for comparison. Random parameters: $\sigma_{sys} = 4$ and $\sigma_{rand} = 5$.

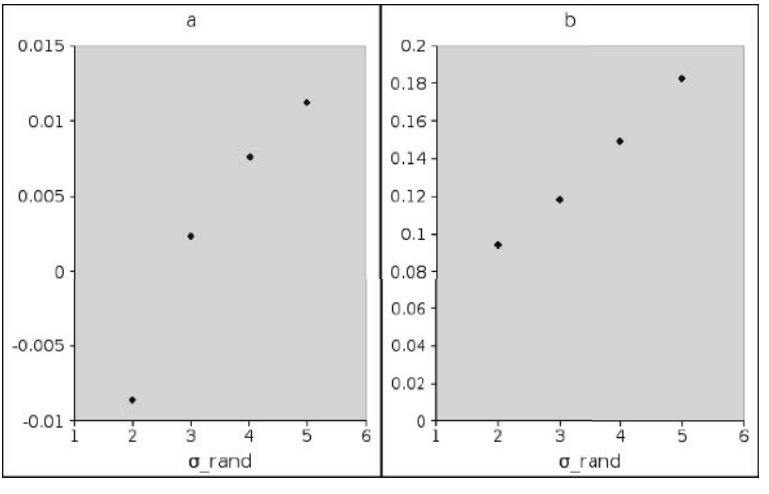

Fig. 4. Fit parameters a and b in simulated data. Both parameters increased with increasing inconsistency of individual classifiers (as parameterized by σ_{rand}).

as random errors in the source labelling as well as the propagation process will tend to cancel each other out. The result will still be subject to systematic bias arising from the input data and the process itself.

We investigated the relationship between the number of classifiers used to form a fused segmentation and the accuracy of the result, and found a model that describes this relation. By performing parameter fitting, it was possible

to quantify the impact of the scale of the transformation used to propagate individual brain segmentations (affine, coarse nonrigid, fine nonrigid). The model thus provides a possible quality measure. In Fig. 2, all the methods of label propagation show increasing precision with increasing n. Thus precision is not in itself a guarantee of appropriate labelling (fusion of many labels without any registration will lead to convergence to a mean label, but not a useful one). It is notable, however, that the approach to consistency with the number of fused labels was slower for less detailed registration methods. The b value was largest when no registration was employed and showed the minimum value when the non-rigid registration with the smallest control point spacing was used. b may therefore represent a means of comparing the accuracy of different registrations.

As Fig. 1 shows, the data from fused label propagation on the human brain was consistent with the model described by Equation 2. For comparisons between independent fused label volumes, the a parameter of the model fit was much smaller, indicating that the averaging that takes place in the label fusion process leads to highly precise (if not necessarily unbiased) results.

The simulations reproduced all key features of the experimental results: the b parameter correlates with σ_{rand} (Fig. 4), showing that it can be used as an indication of the precision of segmentations fused in the experiment. The a parameter describes the accumulated systematic error that cannot be eliminated by considering further classifiers in the fusion process. We conclude that the model (Equation 2) describes the core behavior well.

The appropriateness of the model depends on the assumption that SI values produced for fused labels are normally distributed. Formally, this assumption cannot be correct, since SI has both upper and lower bounds. Nevertheless, SI values produced by random perturbations of the label boundary passed a test of consistency with a normal distribution. This makes the use of the chosen model plausible. Our results suggest that the model can be used to estimate the veracity of fused propagated labels in the absence of a gold standard segmentation. In future work, we are planning to assess the usability of the model for assessing segmentation quality when using fused label propagation in clinical and scientific application scenarios.

References

1. Svarer, C., Madsen, K., Hasselbalch, S.G., Pinborg, L.H., Haugbol, S., Frokjaer, V.G., Holm, S., Paulson, O.B., Knudsen, G.M.: MR-based automatic delineation of volumes of interest in human brain PET images using probability maps. Neuroimage **24**(4) (2005) 969–979
2. Iosifescu, D.V., Shenton, M.E., Warfield, S.K., Kikinis, R., Dengler, J., Jolesz, F.A., Mccarley, R.W.: An automated registration algorithm for measuring MRI subcortical brain structures. Neuroimage **6**(1) (1997) 13–25
3. 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.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron **33**(3) (2002) 341–355

4. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans Pattern Analysis and Machine Intelligence **20**(3) (1998) 226–239
5. Rohlfing, T., Brandt, R., Menzel, R., Maurer, C.R.: Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. Neuroimage **21**(4) (2004) 1428–1442
6. Klein, A., Mensh, B., Ghosh, S., Tourville, J., Hirsch, J.: Mindboggle: Automated brain labeling with multiple atlases. BMC Med Imaging **5**(1) (2005)
7. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging **23**(7) (2004) 903–921
8. Rohlfing, T., Russakoff, D.B., Maurer, C.R.: Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans Med Imaging **23**(8) (2004) 983–994
9. Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S.: Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum Brain Mapp **19**(4) (2003) 224–247
10. Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Berg, J.H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M.: Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage **23 Suppl 1** (2004)
11. Studholme, C., Hill, D.L.G., Hawkes, D.J.: An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition **32**(1) (1999) 71–86
12. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging **18**(8) (1999) 712–721
13. Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C.: Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging **13**(4) (1994)