Multivariate Neuroanalysis with ANTsR: Application to Supervised Brain Segmentation with Concatenated Random Forests

Nicholas J. Tustison[a,1], K. L. Shrinidhi[b], Jeffrey T. Duda[b], Philip A. Cook[b], Christopher Durst[a], James C. Gee[a], Murray C. Grossman[a], Max Wintermark[a], Brian B. Avants[b]

[a]*Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA*
[b]*Penn Image Computing and Science Laboratory, University of Pennsylvania, Philadelphia, PA*

Abstract

Improvements in image acquisition and increasingly sophisticated statistical techniques for neuroanalysis have spurred the development of advanced computational frameworks for studying the brain—many of which have been made publicly available. One such popular toolkit is the open source Advanced Normalization Tools (ANTs) which contains a suite of well-vetted processing tools for core data transformation tasks such as image registration, image segmentation, inhomogeneity field correction and template building. However, despite the numerous solutions afforded by such tools, neuroscience (and data analysis in general) requires a statistical platform for making inferences with respect to given hypotheses once the requisite data transformations have been performed. In this paper, we describe the coupling of ANTs with the widely-used R language—an open source environment for statistical processing and data visualization—which we denote as ANTsR. One of the many benefits from such an integration is the accessibility to advanced statistical and machine learning techniques. To showcase the flexibility and power of ANTsR, we apply the combination of one such technique, Random Forests, and ANTs processing tools to the difficult problem of brain tumor segmentation. This includes evaluation on the public data set from the MICCAI 2012 BRATS challenge consisting of both real and simulated data. To facilitate reproducibility, all scripts and data used in this paper are publicly available for download.

Keywords: advanced normalization tools, BRATS, brain tumor segmentation, R project, random forests

1. Introduction

ANTs (Advanced Normalization Tools) originated with the open source availing of state-of-the-art registration algorithms for neuroimage analysis [3] built upon the mature and well-vetted Insight Toolkit of the National Institutes of Health. Since then, the toolkit has grown to include several algorithmic solutions necessary for robust medical image analysis including bias correction [25], *n*-tissue multivariate segmentation [4], template construction [5], and cortical thickness estimation [11] (many of which have been introduced into ITK partially in an attempted leveraging of Linus's Law).[2] However, in the evolution of the toolkit, it became clear (as neuroimaging research certainly falls under the popular rubric of "big data analysis") that robust statistical machinery was lacking for proper inferences regarding the data produced by the various ANTs tools. ANTsR was developed specifically to provide an interface between a powerful neuroimaging toolkit for producing reliable imaging data transformations and the R project[3] for statistical computing and visualization thus providing a complete set of tools for multivariate image analysis.

In addition to describing ANTsR basics and how it can be generally used in multivariate neuroimaging studies, we showcase its use in a particularly salient application for performing supervised brain segmentation using random forests. Although random forests have been proposed previously in the literature for supervised brain segmentation (e.g. [13, 29]), we demonstrate that ANTsR significantly facilitates the construction of a fully functional, parallelizable workflow for such a task and that performance exceeds that of the top competitors in the recent Multimodal Brain Tumor Segmentation Challenge as part of the MICCAI 2012 conference.[4]

2. Materials and Methods

2.1. ANTsR: An ANTs/R Interface

2.1.1. Installation

The ANTsR package is publicly available on the github project hosting service.[5] Prior to installation of ANTsR, several external R packages need to be installed including: `Rcpp`, `signal`, `timeSeries`, `mFilter`, `doParallel`, `robust`, `magic`, `knitr`, `pixmap`, `rgl`, `misc3d` which is facilitated by the `install.packages()` mechanism. Additionally, in order to perform the supervised brain segmentation as described in later sections, one would need to also install `randomForest`, `snowfall`, `rlecuyer`, and `ggplot2`.

[1]Corresponding author: PO Box 801339, Charlottesville, VA 22908; T: 434-924-7730; email address: ntustison@virginia.edu.

[2]"Given enough eyeballs, all bugs are shallow." –Linus Torvalds

[3]http://www.r-project.org

[4]http://www2.imm.dtu.dk/projects/BRATS2012/

[5]https://github.com/stnava/ANTsR

CMake[6] is an open source tool for the management and building of large-scale software projects. It is used to coordinate the downloading of external packages, such as the Insight Toolkit (ITK)[7] and ANTs. Detailed instructions for download and installation can be found on the ANTsR github website.

2.2. Supervised Brain Segmentation

Given a set of training data consisting of labeled (csf, gray matter, white matter, edema, and tumor), multimodal brain image data, supervised brain segmentation entails the generation of a statistical model from such training data which can then be extended to testing data, i.e. unlabeled brain data. In subsequent sections, we introduce the generic *random forest* modeling framework which takes as input a set of *feature images* (also described in a later section) and labels for each training subject and outputs a statistical model. This model, in turn, can then be used to provide a probabilistic estimate of the labels in an unlabeled subject.

One of the core extensions that we provide in this work is the use of concatenated random forests for improved probabilistic estimation of the labels. As we demonstrate in the Results section, the set of feature images employed work sufficiently well for good performance on the training data (which exceeds that of what has been currently published in the literature). However, we discovered that further improvements could be gained by using the probabilistic label estimates as input to enhance preprocessing before a second round of feature image generation including modality-specific, prior-based *n*-tissue segmentation and random forest model creation.

2.2.1. Random Forests

Several machine learning concepts were integrated to create the random forests framework first articulated in its entirety by Breiman et al. [8] for performing classification/regression. Although decision trees had been previously explored in the literature, it was the success of "boosting"-style machine learning techniques, such as AdaBoost [24, 12], which influenced the aggregation of such decision trees into "forests" with randomized node optimization for improved classification/regression performance [15, 1]. The final element of bootstrap aggregating or "bagging" (i.e. random sampling of the training data) was introduced by Breiman [7] to achieve improved accuracy.

Early adoption [27] and success in the computer vision community has led to a recent surge within the medical image analysis community of using random forests for handling complex classification/regression tasks including normal brain segmentation [28], MS lesion segmentation [13], multimodal brain tumor segmentation [29, 14], brain extraction [16], anatomy detection in computed tomography [9], and segmentation of echocardiographic images [26]. A thorough introduction for those interested in delving deeper into the more theoretical aspects of random forests can be found in [10].

One of the principal advantages of R is the extensive community of developers who have contributed on the order of thousands of packages extending R's capabilities beyond its core functionality. Most relevant for the work described in this paper is the `randomForest` package developed from Breiman's original Fortran code by Liaw and Wiener (described briefly in [18]).

2.2.2. Multi-Modal Feature Image Preprocessing

Although several studies have pointed out the importance of intensity normalization and bias correction, our experience with the training data illustrated a degradation in performance when one or both steps (using [21] and N4 [25], respectively) were performed due to the presence of the tumor/edema complex.

As a corrective, for the first stage we simply windowed the image intensity for all images to be between the quantiles $[0.01, 0.99]$ and subsequently rescaled to $[0, 1]$. From these "corrected" images, the first set of feature images were derived. For the training cohort, these data were used to create the random forest regression model for the first stage. During the second stage, the probabilistic estimates of the white matter and gray matter labels were used to generate a "pure tissue weight mask" to estimate the bias field using N4 (although the resulting bias field estimation was used to correct the image within the entire cerebral mask). Formally, this involved generation of a probabilistic map defined as:

$$P_{pure\ tissue}(\mathbf{x}) = \sum_{i=1}^{N} P_i(\mathbf{x}) \prod_{j=1, j\neq i}^{N} \Big(1 - P_j(\mathbf{x})\Big) \qquad (1)$$

where N is the set of user-selected tissue labels (in our case $N = 2$ consisting of the gray and white matter probability maps).

Both rescaling and weighted bias correction were applied to produce the "corrected images" for the second stage resulting in modified features images for the second stage. Note that we perform a similar iterative scenario for normal brain segmentation [4] (encapsulated in the ANTs script `antsAtroposN4.sh`).

2.2.3. Multi-Modal Feature Image Generation

Key to any supervised regression or classification protocol are the selected features for training and subsequent testing. Based on previous work and our own experience, we selected the following feature images to showcase the supervised segmentation strategy developed in this work.

- Per modality (FLAIR, T1, T1C, T2)
  - First-order neighborhood statistical images: mean, variance, skewness, and entropy. Neighborhood radius = 3.
  - GMM-based posteriors: CSF, gray matter, white matter, edema, and tumor
  - GMM connected component geometry features: volume, volume to surface area ratio, eccentricity, elongation

[6] http://www.cmake.org/
[7] http://www.itk.org/

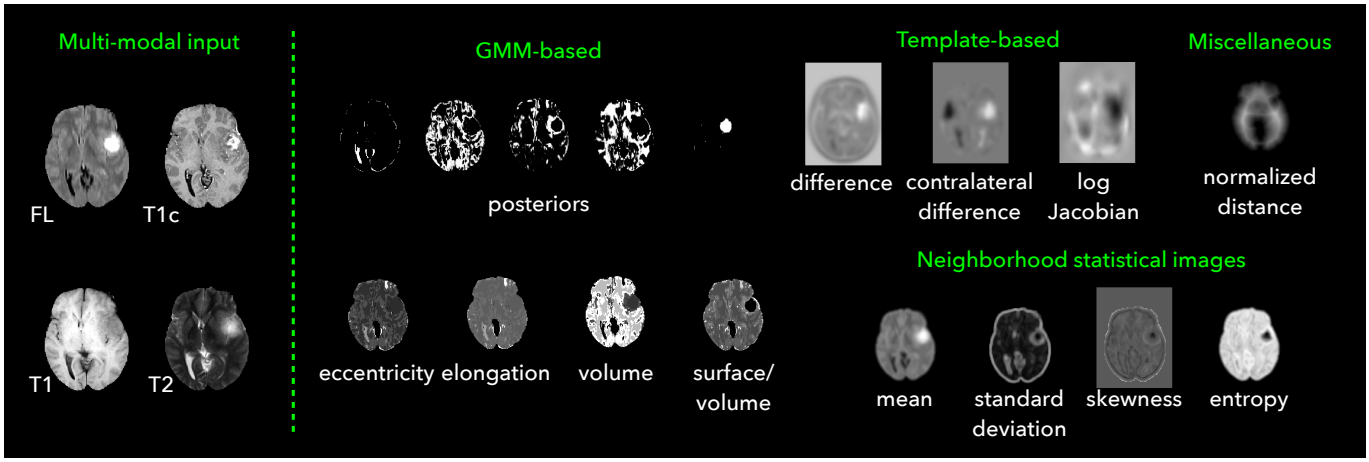

Figure 1: Feature images from the BRATS_HG0004 data set.

  - Template-based: symmetric template difference and contralateral difference. Gaussian smoothing (σ = 4mm).

- Not modality specific
  - Normalized Euclidean distance
  - Log Jacobian
  - T1C,T1 difference image

For each modality, we create four first-order statistical feature images, five Gaussian mixture model (GMM)-based posterior probability feature images, four geometry features generated from the GMM posterior probability images based on connected components, and two difference images using symmetric template construction for a total of 4 modalities $\times$ (4 + 5 + 4 + 2) feature images per modality = 60 total feature images. We employ two additional images consisting of the Euclidean distance image [19] created from the skull-stripped binary mask rescaled to the range $[0, 1]$ and the log Jacobian image derived from the spatial normalization of the symmetric multivariate template and individual subject images. Given the intensity corrected images the corresponding multivariate template images, and a brain mask for each subject creation of all feature images is performed using the script `createFeatureImages.sh`.

Prior cluster centers for specific tissue types learned from training data [23] are used in the GMM to create multiple feature images. Given M tissue types (e.g. CSF, gray matter, white matter, edema, and tumor), a GMM formulates the probability distribution at each voxel, $\mathbf{x}$, as the sum of Gaussian components, $\mathcal{N}(\mathbf{x}|\mu, \sigma)$, i.e.

$$p\left(\mathbf{x}|\mu_m, \sigma_m, \lambda_m\right) = \sum_{i=1}^{M} \lambda_m \mathcal{N}(\mathbf{x}|\mu_m, \sigma_m) \tag{2}$$

where $\sum_{m=1}^{M} \lambda_m = 1$. One popular method for determining the parameters of the GMM is maximum likelihood estimation which can performed using the Atropos segmentation tool [4]. In contrast to previous generative modeling approaches for multi-modal tumor segmentation (e.g. [22, 29]), we do not use multivariate Gaussians to specify tissue probabilities but rather incorporate each univariate probability map into the feature vector of the training data. As pointed out in [20], multivariate modeling might obscure the distinct biological information provided by each modality. Instead, we let the random forest construction process determine the optimal combination of such multivariate information. Additionally, maximum posterior labeling from the GMM processing is used to determine the connected components for each label. Geometric features (assigned voxel-wise) include the physical volumes of each connected component including the volume to surface area ratio, the elongation, and eccentricity.

In order to better characterize deviations from normal multimodal brain shape and appearance, several features were derived using population-specific multivariate template construction. A recent neuroimaging reproducibility study by Landman et al. resulted in an open data cohort of 21 normal individuals, each imaged twice, comprising several modalities including arterial spin labeling, fluid attenuated inversion recovery (FLAIR), diffusion tensor imaging, functional imaging, T1, and T2 [17].

Given K image modality types for N subjects, $\mathbf{I} = \{I_1, I_2, \ldots, I_K\}$, multivariate template construction iterates between optimizing the set of diffeomorphic transforms between the subjects and the template, $\left\{\left(\phi_1, \phi_1^{-1}\right), \ldots, \left(\phi_N, \phi_N^{-1}\right)\right\}$ and constructing the optimal multivariate template appearance $\mathbf{J} = \{J_1, J_2, \ldots, J_K\}$ to minimize the following cost function:

$$\sum_{n=1}^{N} \left[D\left(\psi(\mathbf{x}), \phi_1^n(\mathbf{x}, 1)\right) + \sum_{k=1}^{K} \lambda_k \Pi_k \left(I_k^n\left(\phi_n(\mathbf{x}, 0.5)\right), J_k\left(\phi_n^{-1}(\mathbf{x}, 0.5)\right)\right) \right] \tag{3}$$

where D is the diffeomorphic shape distance,

$$D\left(\phi(\mathbf{x}, 0), \phi(\mathbf{x}, 1)\right) = \int_0^1 \|v(\mathbf{x}, t)\|_L dt \tag{4}$$

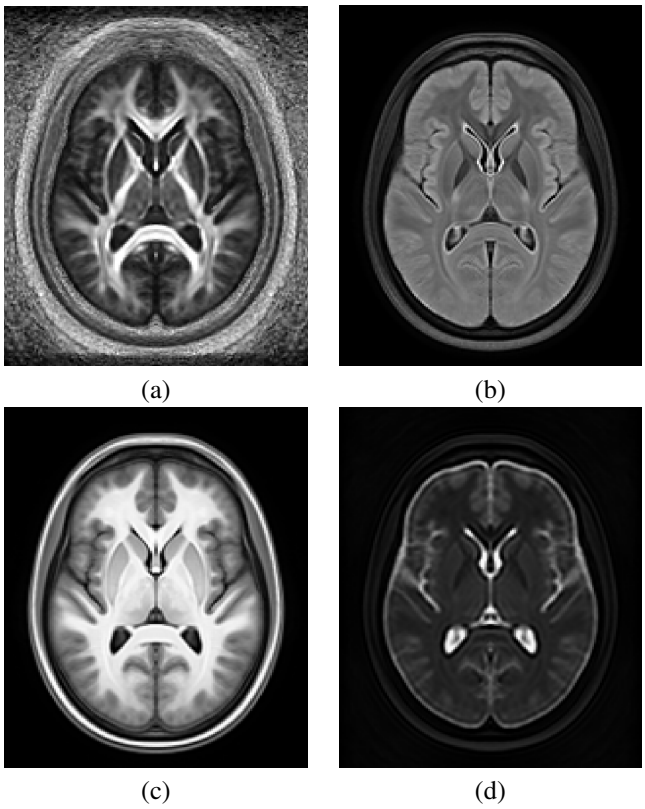

Figure 2: Multivariate symmetric template created from the Kirby 21 data described in [17]. Shown are the (a) fractional anisotropy (FA), (b) FLAIR, (c) MPRAGE, and (d) T2 template components.

dependent on the choice of linear operator, L, and v is the velocity field

$$v\left(\phi(\mathbf{x},t)\right) = \frac{d\phi(\mathbf{x},t)}{dt}, \quad \phi(\mathbf{x},0) = \mathbf{x}. \tag{5}$$

Each pairwise registration employing the similarity metric Π_k can be assigned a relative weighting, λ_k, to weight a particular modality's influence in the construction process. Further theoretical details can be found in [2, 5]. In terms of implementation, this algorithm is encapsulated in the script `antsMultivariateTemplateConstruction.sh`, available in the ANTs repository, which permits parallel processing on an individual workstation or on a large computational cluster.

2.2.4. Brain Tumor Segmentation Challenge Data

The Brain tumor segmentation Associated with the 2012 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI),[8] the

[8] http://www.miccai2012.org

2.2.5. MS lesion segmentation challenge Challenge Data

3. Results

4. Discussion and Conclusions

Acknowledgments

References

[1] Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural Computation 9, 1545–1588.

[2] Avants, B., Duda, J. T., Kim, J., Zhang, H., Pluta, J., Gee, J. C., Whyte, J., Nov 2008. Multivariate analysis of structural and diffusion imaging in traumatic brain injury. Acad Radiol 15 (11), 1360–75.

[3] Avants, B. B., Epstein, C. L., Grossman, M., Gee, J. C., Feb 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 12 (1), 26–41.

[4] Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A., Gee, J. C., Dec 2011. An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics 9 (4), 381–400.

[5] Avants, B. B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., Gee, J. C., Feb 2010. The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49 (3), 2457–66.

[6] Bauer, S., Fejes, T., Slotboom, J., Wiest, R., Nolte, L.-P., Reyes, M., October 2012. Segmentation of brain tumor images based on integrated hierarchical classification and regularization. In: Proceedings of MICCAI-BRATS 2012. pp. 10–13.

[7] Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

[8] Breiman, L., 1996. Random forests. Machine Learning 24 (2), 123–140.

[9] Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., Siddiqui, K., Jan 2013. Regression forests for efficient anatomy detection and localization in computed tomography scans. Med Image Anal.

[10] Criminisi, A., Shotton, J., Konukoglu, E., 2011. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. Tech. rep., Microsoft Resaerch.

[11] Das, S. R., Avants, B. B., Grossman, M., Gee, J. C., Apr 2009. Registration based cortical thickness measurement. Neuroimage 45 (3), 867–79.

[12] Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139.

[13] Geremia, E., Clatz, O., Menze, B. H., Konukoglu, E., Criminisi, A., Ayache, N., Jul 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. Neuroimage 57 (2), 378–90.

[14] Geremia, E., Menze, B. H., Ayache, N., October 2012. Spatial decision forests for glioma segmentation in multi-channel MR images. In: Proceedings of MICCAI-BRATS 2012.

[15] Ho, T. K., 1995. Random decision forests. In: Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on. Vol. 1. pp. 278–282 vol.1.

[16] Iglesias, J. E., Liu, C.-Y., Thompson, P., Tu, Z., 2010. Agreement-based semi-supervised learning for skull stripping. Med Image Comput Comput Assist Interv 13 (Pt 3), 147–54.

[17] Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A. D., Bogovic, J. A., Hua, J., Chen, M., Jarso, S., Smith, S. A., Joel, S., Mori, S., Pekar, J. J., Barker, P. B., Prince, J. L., van Zijl, P. C. M., Feb 2011. Multi-parametric neuroimaging reproducibility: a 3-T resource study. Neuroimage 54 (4), 2854–66.

[18] Liaw, A., Wiener, M., 2002. Classification and regression by randomforest.

[19] Maurer, C. R., Rensheng, Q., Raghavan, V., 2003. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. Pattern Analysis and Machine Intelligence, IEEE Transactions on 25 (2), 265–270.

[20] Menze, B. H., Van Leemput, K., Lashkari, D., Weber, M.-A., Ayache, N., Golland, P., 2010. A generative model for brain tumor segmentation in multi-modal images. Med Image Comput Comput Assist Interv 13 (Pt 2), 151–9.

[21] Nyúl, L. G., Udupa, J. K., Zhang, X., Feb 2000. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging 19 (2), 143–50.

Table 1: Dice scores from the MICCAI 2012 BRATs Study

	High-grade (real)		Low-grade (real)		High-grade (simulated)		Low-grade (simulated)	
Method	Edema	Tumor	Edema	Tumor	Edema	Tumor	Edema	Tumor
[29]	0.70 ± 0.09	0.71 ± 0.24	0.44 ± 0.18	0.62 ± 0.27	0.65 ± 0.27	0.90 ± 0.05	0.55 ± 0.23	0.71 ± 0.20
[6]	0.61 ± 0.15	0.62 ± 0.27	0.35 ± 0.18	0.49 ± 0.26	0.68 ± 0.26	0.90 ± 0.06	0.57 ± 0.24	0.74 ± 0.10
ANTsR	0.65 ± 0.15	0.66 ± 0.28	0.49 ± 0.16	0.65 ± 0.21	0.68 ± 0.25	0.91 ± 0.08	0.61 ± 0.25	0.84 ± 0.09
w/ Atropos	0.68 ± 0.15	0.67 ± 0.30	0.50 ± 0.15	0.67 ± 0.23	0.74 ± 0.26	0.92 ± 0.09	0.65 ± 0.26	0.84 ± 0.08

[22] Prastawa, M., Bullitt, E., Moon, N., Van Leemput, K., Gerig, G., Dec 2003. Automatic brain tumor segmentation by subject specific modification of atlas priors. Acad Radiol 10 (12), 1341–8.

[23] Reynolds, D. A., 2009. Gaussian mixture modeling. In: Li, S. Z., Jain, A. K. (Eds.), Encyclopedia of Biometrics. Springer US, pp. 659–663.

[24] Schapire, R., 1990. The strength of weak learnability. Machine Learning 5, 197–227.

[25] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., Gee, J. C., Jun 2010. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29 (6), 1310–20.

[26] Verhoek, M., Yaqub, M., McManigle, J., Noble, J. A., 2011. Learning optical flow propagation strategies using random forests for fast segmentation in dynamic 2D and 3D echocardiography. In: Suzuki, K., Wang, F., Shen, D., Yan, P. (Eds.), Machine Learning in Medical Imaging. Vol. 7009 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 75–82.

[27] Viola, P., Jones, M., Snow, D., 2005. Detecting pedestrians using patterns of motion and appearance. International Journal of Computer Vision 63, 153–161.

[28] Yi, Z., Criminisi, A., Shotton, J., Blake, A., 2009. Discriminative, semantic segmentation of brain tissue in MR images. Med Image Comput Comput Assist Interv 12 (Pt 2), 558–65.

[29] Zikic, D., Glocker, B., Konukoglu, E., Shotton, J., Criminisi, A., Ye, D. H., Demiralp, C., Thomas, O. M., Das, T., Jena, R., Price, S. J., October 2012. Context-sensitive classification forests for segmentation of brain tumor tissues. In: Proceedings of MICCAI-BRATS 2012. pp. 1–9.