A ROBUST AND SENSITIVE VOXEL-BASED METHOD FOR MEASURING CORTICAL THICKNESS CHANGE USING FUZZY CORRESPONDENCE

Saurabh Garg, Anthony Traboulsee, David K. B. Li and Roger Tam

MS/MRI Research Group, University of British Columbia, Vancouver, Canada
email: saurabh@msmri.medicine.ubc.ca

Abstract: Cortical thickness is an important surrogate biomarker for evaluating the progression of neurodegenerative diseases. We propose a new method for measuring cortical thickness changes on longitudinal magnetic resonance images (MRIs). The method is voxel-based for computational efficiency and sensitivity to subtle changes, but aims for robustness in establishing correspondences by using intuitive features that are defined on a cortical skeleton computed in each scan. In contrast to existing longitudinal methods, the proposed method does not require deformable registration. Instead, we use cortex specific matches and fuzzy correspondence, which allows a skeletal point in one scan to be partially matched to multiple points in another scan, thereby enhancing the stability of the matches. Our experiments show that the proposed method is comparable in scan-rescan reproducibility with a state-of-the-art surface-based method, and demonstrates greater sensitivity to clinically relevant changes on a dataset containing MRIs of 100 secondary progressive multiple sclerosis subjects.
Keywords: Cortical Thickness, Atrophy, Longitudinal Measurement.

Introduction

Thinning of the cortex has been linked to various neurological disorders such as multiple sclerosis (MS). A robust and sensitive method for measuring longitudinal changes in cortical thickness using magnetic resonance imaging (MRI) is highly desirable. However, there are several key challenges involved. The typical acquired MRI resolution for cortical analysis (1 mm^3) is coarse relative to the mean cortical thickness (2.5 mm [1]), resulting in a large partial volume effect. In addition, the annual thickness change is small (~1% [2]) for secondary progressive MS (SPMS).

A number of approaches have been proposed for cortical thickness measurement, and most aim to achieve a balance between robustness, typically achieved through some form of regularization, and sensitivity to real change. Surface-based methods (e.g., [2,3]), which typically create two triangulated meshes for the inner and outer cortical surfaces of each scan, are a relatively successful class of methods due to their ability to enforce certain constraints such as smoothness and topology during surface reconstruction, which reduces the impact of noise and minor segmentation errors. However, surface reconstruction is computationally very expensive, with current methods taking 12 to 30 hours to process a subject [2,3]. In contrast, voxel-based methods, which can perform measurements directly on the grey matter (GM) segmentation, are generally much faster. However, voxel-based methods do not have the benefit of a regularizing surface, and therefore can be less accurate and reproducible than surface-based methods [1].

A number of methods have been proposed specifically for processing longitudinal data, with the idea that using multiple scans across time would increase the signal-to-noise ratio in unchanged areas, while the real changes would be detected with greater sensitivity. Longitudinal processing requires establishing anatomical correspondence across time, and most current methods use deformable registration for this purpose (e.g., [3,4]). While longitudinal methods have shown advantages over otherwise comparable cross-sectional methods, deformable registration has its disadvantages in that the choice of registration method and parameters can be confounding factors.

We propose a new voxel-based method called LCT (Longitudinal Cortical Thickness), for measuring longitudinal changes in cortical thickness between pairs of scans. Rather than a general alignment approach like deformable registration, the proposed method only matches points in the cortex using three intuitive positional and shape features that are defined on a cortical skeleton computed on each scan. To improve the stability of the matches, we use fuzzy correspondence in which each point in one scan can partially match several points in the other scan. After computing the matches between the two skeletons, the thickness measurements are performed by integrating GM probabilities along common directional vectors computed using both scans. We validate the method by applying it to a small reproducibility dataset and a large clinical dataset and comparing the global and lobe-level thickness results with those produced with a state-of-the-art method, FreeSurfer [3].

Materials and methods

To evaluate the proposed method, we applied it to two longitudinal datasets of 3D T1-weighted brain MRIs: 1) images from 15 subjects with two scans each, acquired an hour apart, to measure scan-rescan reproducibility, 2) images from a completed multi-center clinical trial in SPMS, consisting of 100

randomly selected subjects, each with two scans acquired two years apart, to measure sensitivity to real changes and their correlations to clinical parameters. The voxel sizes have ranges s_x = 0.99 to 1.20 mm, s_y = 0.98 to 1.24 mm, and s_z = 0.93 to 1.24 mm, but all scans were resampled to have 1 mm isotropic voxels as part of preprocessing. The mean age of the subjects was 49.5 years. The MS functional composite (MSFC), a widely used MS clinical score composed of 3 subscores: 25-Foot walk (T25W), Paced Auditory Serial Addition Test (PASAT) and 9-Hole Peg Test (9-HPT), was used to test the clinical relevance of the results.

Proposed method

A number of preprocessing steps are applied before the thickness measurement. First, the non-parametric non-uniformity normalization (N3 [5]) algorithm is used for inhomogeneity correction in each image separately. Each pair of scans are then rigidly registered to each other using symmetric transformations to a midpoint, then a paired inhomogeneity correction [6] is applied to minimize the remaining differences in the bias fields.

Next, skull-stripping is performed using two methods: the Brain Extraction Tool (BET [7]) and the skull-stripping program in FreeSurfer (mri_watershed). We found that both occasionally leave extraneous tissue, and that intersecting the two masks largely resolves this problem. Next, cerebellum exclusion is performed on each scan by affinely registering to an atlas with a manually segmented cerebellum and transferring the label onto each scan. The brain images are segmented using an expectation maximization and hidden Markov random field approach (FAST [8]) to obtain the probabilistic classification of GM. The non-cortical GM is then removed by processing each axial slice independently and keeping the largest connected component near the skull and rejecting any smaller disconnected components within.

Thickness change computation – The core of the algorithm consists of the following steps: 1) compute a skeletal representation of the cortical GM in each scan; 2) compute fuzzy correspondences between the scans on the skeletal points; 3) for each pair of matched points, average the two 3D normals to the skeletons to compute a common normal; 4) integrate the GM probabilities along the common normal in both scans, weighted by the strength of the match, to compute the thickness in each scan; 5) compute the difference in thickness.

The skeletal line on each scan is obtained by 2D morphological binary thinning on the GM segmentation that has been thresholded at 90%. Next, for every point on the skeletal line in the first scan, a set of the closest matches are found on the skeletal line in the second scan based on three positional and shape features: spatial coordinates, unit normal direction, and shape context [9]. Shape context can be intuitively thought of as a spatial histogram that captures the distribution of neighboring points relative to a reference point.

Let the skeletal point at a time point *t* and position *p* be $\boldsymbol{s}_p^{(t)}$. Between points $\boldsymbol{s}_p^{(1)}$ and $\boldsymbol{s}_q^{(2)}$, the cost of the mismatch in positional coordinates is calculated as:

$$c_1\left(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)}\right) = \frac{1}{K_{\mathrm{pc}}} \sum_{j \in x,y,z} \left(P_j\left(\boldsymbol{s}_p^{(1)}\right) - P_j\left(\boldsymbol{s}_q^{(2)}\right)\right)^2 \quad (1)$$

where $P_j(\boldsymbol{s})$ is the value of the j^{th} component of $\boldsymbol{s}$ and K_{pc} is a normalization constant. Similarly, for unit normals, the cost of the mismatch is calculated as:

$$c_2\left(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)}\right) = 1 - \boldsymbol{N}\left(\boldsymbol{s}_p^{(1)}\right).\boldsymbol{N}\left(\boldsymbol{s}_q^{(2)}\right) \quad (2)$$

where $\boldsymbol{N}(\boldsymbol{s})$ is the unit 3D normal at point $\boldsymbol{s}$ in a square window of size 5 and (.) represents the dot product of the vectors. Finally, the cost of the mismatch in shape context is calculated using a χ^2 statistic:

$$c_3\left(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)}\right) = \frac{1}{K_{\mathrm{sc}}} \sum_{i=1}^{5} \sum_{j=1}^{12} \frac{\left(V_{ij}\left(\boldsymbol{s}_p^{(1)}\right) - V_{ij}\left(\boldsymbol{s}_q^{(2)}\right)\right)^2}{V_{ij}\left(\boldsymbol{s}_p^{(1)}\right) + V_{ij}\left(\boldsymbol{s}_q^{(2)}\right)} \quad (3)$$

where $V_{ij}(\boldsymbol{s})$ is the value of the log-polar histogram at point $\boldsymbol{s}$ and K_{sc} is a normalization constant. To construct the histogram, a circular window of radius 4 is partitioned into 5 radial bins and 12 angular θ bins.

The normalization constants K_{pc} and K_{sc} are each set to the maximum individual cost computed for that feature, so that $c_i \in [0,1]$. The individual cost from each feature is weighted differently towards the final cost:

$$c\left(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)}\right) = \sum_{i=1}^{3} w_i c_i\left(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)}\right) \quad (4)$$

where w_i is the weight of feature *i* between points $\boldsymbol{s}_p^{(1)}$ and $\boldsymbol{s}_q^{(2)}$. We determined the weights w_i empirically by applying pseudo-random deformations to a scan with a GM segmentation, measuring the change in GM volume using Jacobian integration (JI), then finding the set of weights in our method that maximizes the agreement of our thickness measurements to the JI. The final weights were constrained to not differ by more than 0.25 from each other for generalizability, resulting in the values 0.25, 0.50, 0.25 for w_1, w_2, w_3, respectively.

Using the computed costs, for each point in the first scan, a maximum of the three closest matches are found in the second scan. In order to constrain the maximum distance of the match, points are compared in a 3D neighborhood window of size 7, which should be sufficiently large to contain the true matches for most MS studies up to 5 years. A match is discarded if its cost c is above a precomputed threshold, which we determined empirically by varying the threshold in an independent set of development scans, and finding the value that maximized the number of matches while producing stable cortical change measurements. The thickness for each time point is then calculated as the sum of GM probabilities along the direction of the average normal ($\boldsymbol{N}_{\mu}$) computed between the unit normals of the two scans. The average normal is used as a method of longitudinal regularization to reduce the variability of normal computations. From the point $\boldsymbol{s}_p^{(1)}$, the GM probabilities are integrated in both forward and backward directions along $\boldsymbol{N}_{\mu}$ to compute the thickness

$$T\left(\boldsymbol{s}_p^{(1)}\right) = \sum_{i=0}^{n_{\mathrm{f}}} p\left(\boldsymbol{s}_p^{(t)} + i\Delta \boldsymbol{s}\boldsymbol{N}_{\mu}\right) + \sum_{i=1}^{n_{\mathrm{b}}} p\left(\boldsymbol{s}_p^{(t)} - i\Delta \boldsymbol{s}\boldsymbol{N}_{\mu}\right) \quad (5)$$

where *p(s)* is the linearly interpolated probability of point $\boldsymbol{s}$ belonging to GM, $\Delta\mathbf{s}$ is the step size (0.25 mm) and n_{f}, n_{b} are the numbers of steps taken in the forward and backward directions until one of the following stopping criteria is met: 1) if there is a break in the expected monotonic decrease in GM probability; 2) if a clinical prior of 4 mm away from the skeleton point in either direction is reached. The total prior of 8 mm is larger than used in some previous work (5 mm in [4]) because we are working directly with the probabilistic segmentation and not a thinner binary segmentation. Next, in order to detect buried sulci (deep, thin sulci where the cerebrospinal fluid, or CSF, is misclassified as GM), each average normal is analyzed to determine whether both the forward and backward directions terminate at the same type of boundary (GM-CSF or GM-WM). In the cases where they do, we take the thickness value at that point to be half of the measured value, similar to how buried sulci is handled in other voxel-based methods (e.g. [10]). Then, the mean change in thickness between matched points, weighted by the strength of each match, is calculated as:

$$\overline{\Delta T} = \frac{1}{P^{(1)}+P^{(2)}} \sum_{p=1}^{P^{(1)}} \sum_{q\in\mathcal{L}_p^{(2)}} W(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)}) \left[T(\boldsymbol{s}_q^{(2)}) - T(\boldsymbol{s}_p^{(1)})\right] \quad (6)$$

where $P^{(t)}$ is the number of skeletal points at time point t, $\mathcal{L}_p^{(2)}$ is the set of points in the second scan that have been matched with $\boldsymbol{s}_p^{(1)}$ and $W\left(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)}\right)$ is the strength of the match between $\boldsymbol{s}_p^{(1)}$ and $\boldsymbol{s}_q^{(2)}$, defined as:

$$W\left(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)}\right) = \frac{S(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)})}{\sum_{q\in\mathcal{L}_p^{(2)}} S\left(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)}\right) + \varepsilon} \quad (7)$$

where $S\left(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)}\right) = c(\boldsymbol{s}_p^{(1)}, \boldsymbol{s}_q^{(2)})^{-1}$ and ε is a small number added for numerical stability. In order to make the measurements symmetric between the two time points, the thickness changes are also calculated in the reverse direction using fuzzy correspondence and the resulting two measurements are averaged.

Experiments and Results

We compared LCT with the longitudinal pipeline of FreeSurfer v5.2.0. It took approximately 2.5 hours for LCT to process a subject, whereas FreeSurfer took about 26 hours on an Intel 2.33 GHz machine with 8GB of RAM. We performed comparisons in global thickness values for both datasets and at the lobe level for MS, with the lobes defined from the FreeSurfer labels and transformed into voxel space. Sensitivity to real change over time, reproducibility, and ability to reveal clinical relevant changes were evaluated for both methods.

Scan-rescan reproducibility – When applied to the 15 pairs of scans acquired an hour apart, LCT computed a mean change of -0.094% (SD = 0.525%) whereas FreeSurfer computed a mean change of 0.007% (SD = 0.452%). Neither mean is statistically different from 0, when one-tailed *t*-tests are applied. The two SDs are comparable, indicating that the methods produced similar scan-rescan variability in this dataset.

Results on the MS dataset over two years – Table 1 shows the global and regional mean thickness change results for the MS dataset with 100 subjects. Our method measured a mean change of -0.242% (SD = 1.398%) over the two-year interval, compared to FreeSurfer, which measured a mean change of -0.561% (SD = 3.867%). One tailed *t*-tests produced similar p-values, 0.092 and 0.150, for LCT and FreeSurfer, respectively, for these global changes. For the regional analysis, LCT seemed slightly more sensitive to changes in the left frontal and occipital lobes than FreeSurfer, producing somewhat lower p-values. It is also notable that LCT provided better left-right symmetry, while FreeSurfer produced some very large SDs in the right hemisphere. Our results are consistent with earlier MS studies that also reported significant changes in cortical thickness in the frontal lobe. Overall, our method and FreeSurfer agree reasonably well, considering that they are two very different approaches. Cross-sectionally, the global mean thickness measurements produced by our method were 2.98 mm (SD = 0.21) and 2.97 mm (SD = 0.22) for Time Points 1 and 2, respectively, while FreeSurfer produced close mean measurements of 3.13 mm (SD = 0.21) and 3.12 mm (SD = 0.22). We evaluated the longitudinal agreement between the methods by computing the correlations between the changes measured (Table 2). The global cross-sectional results correlate moderately well (~0.6), while the

Table 1: Mean cortical thickness change (%) over 2 years computed by LCT and FreeSurfer on 100 SPMS subjects. Our method produced lower mean measurements for all regions, but lower SDs as well. The statistical tests show slightly greater sensitivity to change for our method. The ** indicates where a p-value was significant ($p < 0.05$) and * indicates a trend towards significance ($0.05 < p < 0.1$).

	Global		Frontal	Temporal	Parietal	Occipital
FS Mean % Change	-0.561	Left Hemi.	**-0.650***	-0.360	-0.420	-0.240
SD	3.867		3.520	3.840	3.310	4.130
LCT Mean % Change	**-0.242***		**-0.350****	-0.010	-0.030	**-0.520****
SD	1.398		1.740	1.700	1.600	2.400
FS Mean % Change		Right Hemi.	-0.490	-1.160	-1.520	-1.080
SD			3.690	9.760	13.260	10.300
LCT Mean % Change			-0.290	-0.040	-0.200	-0.090
SD			1.750	1.660	1.680	2.240

Table 2: Pearson correlation coefficient (r) between thickness measurements of FreeSurfer and our method over 2 years computed on the MRIs of 100 SPMS subjects. All the p-values from the correlations produced are statistically significant ($p < 0.05$) except for two which are marked with [a]. Overall, the proposed method and FreeSurfer agree reasonably well.

	Time Point 1		**Time Point 2**		**% Thickness change**	
Global	0.581		0.594		0.242	
	Left	Right	Left	Right	Left	Right
Frontal	0.752	0.810	0.660	0.718	0.450	0.346
Temporal	0.531	0.578	0.461	0.340	0.177	0.057[a]
Parietal	0.705	0.752	0.701	0.580	0.446	0.331
Occipital	0.547	0.528	0.523	0.471	0.061[a]	0.220

Table 3: Pearson correlation coefficient (r) between MS clinical scores and % change in thickness measurements over 2 years, as computed in 100 SPMS subjects by FreeSurfer (FS) and our method (LCT). LCT was able to measure thickness changes that correlate significantly with T25W at the global level, and with PASAT in the frontal and temporal lobes, while FS did not. The ** indicates where the p-value was significant ($p < 0.05$).

		Global	Left Hemisphere				Right Hemisphere			
			Frontal	Temporal	Parietal	Occipital	Frontal	Temporal	Parietal	Occipital
LCT	MSFC	-0.004	-0.019	-0.012	0.089	0.032	-0.027	0.119	0.058	0.032
	PASAT	-0.013	**0.241****	**0.213****	0.128	0.109	**0.272****	**0.206****	0.117	0.063
	T25W	**0.208****	0.052	0.070	0.021	0.078	0.093	0.121	0.070	0.137
	9-HPT	0.136	-0.087	-0.154	-0.078	-0.135	-0.101	-0.148	0.008	-0.146
FS	MSFC	0.006	-0.024	-0.019	-0.039	0.028	-0.005	0.041	0.017	0.022
	PASAT	0.077	0.056	0.072	0.146	0.148	0.075	-0.009	0.031	0.040
	T25W	0.005	0.075	0.005	0.082	0.003	-0.003	0.001	-0.010	-0.093
	9-HPT	0.027	-0.023	-0.041	-0.013	-0.016	-0.048	0.066	0.111	0.140

correlations for the regional cross-sectional measurements range from about 0.3 to 0.8, with the strongest agreement in the frontal and parietal lobes. In the longitudinal results, mostly moderate correlations between the two methods are seen, with the global changes having a correlation of 0.241 and the regional changes having correlations ranging from about 0.2 to 0.4, again with the frontal and parietal lobes exhibiting the strongest agreement between the two methods.

To evaluate the clinical relevance of the results from the two methods, correlations were also computed between the percentage changes in thickness and percentage changes in clinical scores (Table 3). The global percentage thickness change computed by our method correlates significantly with the percentage change in the T25W scores ($r = 0.208$, $p < 0.05$). In addition, the regional analysis revealed significant correlations of percentage change in thickness measured with LCT with percentage change in the PASAT in the frontal and temporal lobes in both hemispheres. The Pearson correlation coefficients in the frontal lobe are slightly higher (0.241 for left and 0.272 for right), compared to the temporal lobe (0.213 for left and 0.206 for right). LCT was able to find significant correlations between longitudinal changes in cortical thickness and clinical scores, which is of strong clinical interest. This positive finding has very rarely been reported in the literature, and was not seen in the FreeSurfer results.

Discussion and Conclusion

We have proposed a new voxel-based longitudinal method for measuring changes in cortical thickness between pairs of MRIs. Unlike other methods, the proposed method does not require deformable registration, but instead uses a robust feature-matching approach specifically targeting the cortex. Fuzzy correspondence is used to enhance the stability of matches. Tests using scan-rescan and clinical datasets show that LCT demonstrates greater sensitivity to clinically relevant changes, as assessed by correlations between thickness changes and changes in clinical scores, than FreeSurfer, while retaining the same reproducibility. Overall, LCT and FreeSurfer agree reasonably well for cross-sectional measurements, but differ longitudinally, which is not unexpected given that LCT takes a very different approach to directly measure thickness changes, while FreeSurfer is still largely a cross-sectional method with affine registration added to reduce longitudinal variability. While the lack of ground truth makes resolving differences difficult, LCT appears to have some advantages in our experiments. Additionally, the relatively fast processing speed makes LCT very practical for large clinical studies.

Acknowledgements

This work was supported by grant funding from NSERC, MS Society of Canada and the Milan and Maureen Ilich Foundation.

References

[1] Clarkson, M.J., Cardoso, M.J., Ridgway, G. R., Modat, M., et al.: A comparison of voxel and surface-based cortical thickness estimation methods. NeuroImage. 2011; 57(3): p. 856-865.

[2] Nakamura, K., Fox, R., Fisher, E. CLADA: Cortical longitudinal atrophy detection algorithm. NeuroImage, 2011; 54(1): p. 278-289.

[3] Reuter, M., Schmansky, N.J., et al. Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage, 2012; 61(4): p. 1402-1418.

[4] Das, S.R., Avants, B.B., Grossman, M., Gee, J.C.: Registration-based cortical thickness measurement. NeuroImage, 2009; 45(3): p. 867-879.

[5] Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging, 1998; 17(1): p. 87-97.

[6] Lewis, E.B., Fox, N.C.: Correction of differential intensity inhomogeneity in longitudinal MR images. NeuroImage, 2004; 23(1): p. 75-83.

[7] Smith, S.: Fast robust automated brain extraction. Human Brain Mapping, 2002; 17(3): p. 143-155.

[8] Zhang, Y., Brady, M., et al.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imag. 2001; 20: p. 45-57.

[9] Belongie, S., Malik, J., et al.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 2002; 24(4): p. 509-522.

[10] Scott, M., Bromiley, P., Thacker, N., Hutchinson, C., Jackson, A.: A fast, model-independent method for cerebral cortical thickness estimation using MRI. Medical image analysis. 2009; 13 (2): p. 269-285.