A SENSITIVE AND EFFICIENT METHOD FOR MEASURING CHANGE IN CORTICAL THICKNESS USING FUZZY CORRESPONDENCE IN ALZHEIMER'S DISEASE

Saurabh Garg *Lisa Tang* *Anthony Traboulsee* *Roger Tam*
for the Alzheimer's Disease Neuroimaging Initiative*

MS/MRI Research Group, University of British Columbia, Vancouver, Canada

ABSTRACT

We present 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 geometric features that are defined on the cortical skeletons of each scan. In contrast to existing longitudinal methods, our method does not require deformable registration but rather performs cortex-specific, feature-based matching in a confidence-weighted manner, 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. Based on comprehensive experiments and statistical analyses on two datasets, our results show that the proposed method demonstrates greater sensitivity to clinically relevant changes than three other state-of-the-art methods and comparable reproducibility.

Index Terms— Cortical Thickness, gray matter, atrophy, longitudinal measurement

1. INTRODUCTION

Thinning of the cortex has been linked to various neurological disorders such as Alzheimer's disease (AD). A robust and sensitive method for measuring changes in cortical thickness using magnetic resonance imaging (MRI) is therefore highly desirable. Several approaches have been proposed for measuring cortical thickness, and most aim to achieve a balance between *reliability* and *sensitivity* to real change [1]. Surface-based methods (e.g., [2, 3]), which typically create two triangulated meshes for the inner and outer cortical surfaces of each scan, achieve reliability by enforcing constraints like smoothness and topology-preservation during surface reconstruction, which nonetheless is very computationally expensive (25-30 hours [3]). In contrast, voxel-based methods, which can perform measurements directly on the gray matter (GM) segmentation, are generally much faster [4, 5] but less accurate and reliable than surface-based methods [6]. To increase sensitivity to real change, a number of segmentation-based cross-sectional methods, which analyze each serial MRI independently [2], have been extended for processing longitudinal data, with the idea that processing multiple scans acquired across time would increase the signal-to-noise ratio in unchanged areas [3]. This approach however requires establishing anatomical correspondence across time via deformable registration [4], which is also very computationally expensive. Overall, the aforementioned methods all achieve reliability and sensitivity at the expense of large resource consumption (both time and memory), making large scale analyses difficult.

In this paper, we propose a longitudinal method Longitudinal Cortical Thickness (LCT) that bypasses the need of dense deformable registration for computational efficiency without compromising reliability and sensitivity. We do so by performing feature-based point-correspondences *only within* the cortical regions. On a high level, it uses probabilistic GM segmentations and derives point-correspondence between cortical skeleton lines extracted from the reference and follow-up scan. Rather than using raw intensities as done in [4, 5], we employ shape-context, and other surface-based features for more accurate matching. Further, we perform matching in a confidence-weighted manner to account for partial volume effects and derive the final correspondences such that the thickness measures are spatio-temporally regularized. The closest work to our proposed method is the minimum line integral (MLI) [7], which also operates on skeletal lines of GM segmentation and uses information from the probabilistic segmentation to estimate cortical thickness. However, unlike our method, MLI does not establish correspondences longitudinally, which makes it less sensitive to subtle changes, as shown in section 3.

We have previously presented an early proof-of-concept in [8]. However, [8] relied partially on 2D features for matching and only provided preliminary results on an in-house dataset. In this work, we extended [8] substantially by 1) computing all features in 3D and 2) by handling buried sulci (deep, thin sulci where the cerebrospinal fluid, or CSF, is

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

often mislabeled as GM; see Figure 1) via a novel step that uses temporal information to account for errors in the GM segmentation. We also performed extensive experiments that evaluated: 1) *reproducibility* via scan-rescan data, and 2) global and regional *sensitivity* in detecting cortical thickness changes, and comparative analyses with 3 other state-of-the-art methods (FreeSurfer [3], DiReCT [4], and MLI [7]).

2. PROPOSED METHOD

Our proposed algorithm takes probabilistic segmentations of the cortical GM as input and performs the following steps: 1) compute a skeletal representation of the cortical GM in each scan; 2) compute fuzzy correspondences between the skeletal points of the two scans; 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; 5) compute the difference in thickness. The skeletal line on each scan is obtained by performing morphological binary thinning on the binary GM segmentation. 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 features: positional coordinates, unit normal direction, and shape context [9].

Let a skeletal point at a time point t be $\mathbf{s}_p^{(t)}$ where p denotes the index. The cost of the mismatch (C) between points $\mathbf{s}_p^{(1)}$ and $\mathbf{s}_q^{(2)}$ is calculated as a weighted sum of 1) the Euclidean distance between their positional coordinates, 2) the angle between their unit normals, and 3) the χ^2 statistic between their shape context descriptor as proposed in [9]. Intuitively, shape context can be thought of as a spatial histogram that captures the distribution of neighboring points relative to a reference point. To construct the histogram, a spherical window of radius 4 is divided into 5 radial, 12 polar and 12 azimuthal bins.

Each of the three individual distances is weighted differently towards the final cost (C). The weight of each feature is determined empirically as explained in [8]. Using the computed costs, for each point in the first scan, 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. The thickness for each time point is then calculated as the sum of GM probabilities along the direction of the average normal ($\mathbf{N}_\mu$) computed between the two unit normals at the matched skeletal points. The average normal is used as a method of temporal regularization to reduce the variability of normal computations. From the point $\mathbf{s}_p^{(1)}$, the GM probabilities are integrated in both forward and backward directions along $\mathbf{N}_\mu$ to compute the thickness $T(\mathbf{s}_p^{(1)})$:

$$T(\mathbf{s}_p^{(1)}) = \sum_{i=0}^{n_f} p(\mathbf{s}_p^{(1)} + i\Delta s\mathbf{N}_\mu) + \sum_{i=1}^{n_b} p(\mathbf{s}_p^{(1)} - i\Delta s\mathbf{N}_\mu) \quad (1)$$

where $p(\mathbf{s})$ is the linearly interpolated probability of point $\mathbf{s}$ belonging to GM, Δs 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 monotonic decrease in GM probability; 2) if a clinical prior of 4 mm away from the skeleton point 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 thresholded and therefore thinner binary segmentation.

Finally, we compute the mean change in thickness between matched points($\mathbf{s}_p^{(1)}, \mathbf{s}_q^{(2)}$), weighted by the strength of each match as:

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

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 $\mathbf{s}_p^{(2)}$ and $W(\mathbf{s}_p^{(1)}, \mathbf{s}_q^{(2)})$ is the strength of the match between $\mathbf{s}_p^{(1)}$ and $\mathbf{s}_q^{(2)}$, defined as:

$$W(\mathbf{s}_p^{(1)}, \mathbf{s}_q^{(2)}) = \frac{S(\mathbf{s}_p^{(1)}, \mathbf{s}_q^{(2)})}{\sum_{q \in \mathcal{L}_p^{(2)}} S(\mathbf{s}_p^{(1)}, \mathbf{s}_q^{(2)})} \quad (3)$$

where $S(\mathbf{s}_p^{(1)}, \mathbf{s}_q^{(2)}) = (c(\mathbf{s}_p^{(1)}, \mathbf{s}_q^{(2)}) + \epsilon)^{-1}$ 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 and the resulting two measurements are averaged.

Handling of buried sulci: Deep, thin sulci are locations where CSF is often mislabeled as GM due to partial volume effect, resulting in overestimation of cortical thickness. Our proposed method addresses this problem with a novel 2-step procedure that leverages temporal information to avoid such overestimation. Firstly, we identify skeletal points that are buried by analyzing the forward and backward directions along the average normal ($\mathbf{N}_\mu$) to determine whether they terminate at the same type of boundary (GM-CSF or GM-WM), in which case they are marked as being in buried sulci. The next step involves correcting the cortical thickness measurements at these locations. There can be two scenarios at each identified sulcal region location: 1) it is buried in *both* the reference and follow-up scans, and 2) it is only buried in one scan (usually the reference scan). The latter case would cause a large difference in thickness measure if left unchecked. In the first case, we take the thickness value at the skeletal point to be half of the measured value as done in previous methods (e.g. [10]). In the second scenario, temporal information is used to correct the mislabeled CSF of one scan by modifying its class probabilities using information from the

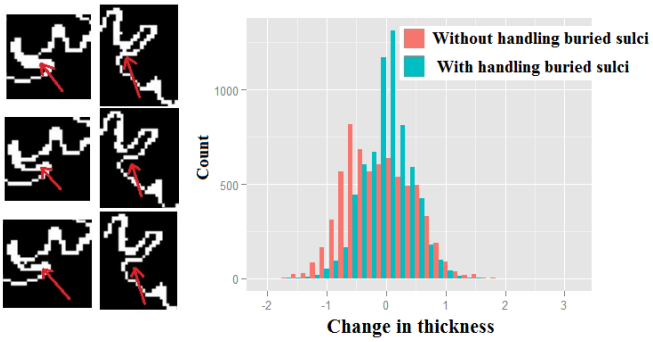

Fig. 1: Selected regions showing the recovered buried sulci. The first row shows two selected regions from a reference scan, the second row shows the same regions on a follow-up scan and the last row shows the recovered buried sulci in the reference scan using temporal information. The histogram shows the distribution of the calculated change in thickness (in mm) at the buried sulcus locations. We can see that not handling the buried sulcus may lead to overestimation of thickness as shown by the histogram centers.

other scan. More specifically, if a buried sulcus in the first scan corresponds to an open sulcus in the second scan, the areas of high CSF confidence (≥ 0.5) in the second sulcus are used to "open up" [4] the first sulcus by increasing CSF probabilities in the corresponding voxels. Figure 1 shows two examples of recovered sulci using the above method. Also, the histogram shows that when buried sulci are not handled, it leads to overestimation of thickness.

3. EXPERIMENTS AND RESULTS

We evaluated our method based on two criteria: 1) scan-rescan reproducibility, 2) sensitivity to change. High sensitivity is desirable as it ensures that our method can detect subtle signals, and high reproducibility ensures we can do so in a consistent manner. We employed two longitudinal datasets of 3D T1-weighted brain MRIs: 1) a dataset of 16 healthy subjects, each with two scans acquired an hour apart, to measure scan-rescan reproducibility, 2) images from 50 AD subjects and 50 age-matched normals randomly selected from the ADNI study [11], each with two scans acquired one year apart, to test each method's ability to distinguish between AD and normal. Using the above criteria, we evaluated 3 other state-of-the-art cortical thickness measurement methods: FreeSurfer v5.2.0, MLI and DiReCT (ANTS 2.1v4).

Scan-rescan reproducibility: Table 1 summarizes the results of scan-rescan reproducibility applied to the 16 pairs of scan-rescans. FreeSurfer produced the smallest mean change of 0.007% and our method produced the second smallest change of -0.074%. The largest mean change was measured by MLI followed by DiReCT but none of the means are statistically different from 0, when 2-tailed t-tests are applied. In addition, the SDs of our method, FreeSurfer, and MLI are comparable, indicating that the methods produced similar scan-rescan variability. On the other hand, DiReCT measurements produced a large SD and a large mean change, suggesting weaker scan-rescan reproducibility.

Table 1: Scan-rescan reproducibility. One sample t-test shows that none of the means are statistically different from 0. The SD indicates that all methods except DiReCT produced similar scan-rescan variability.

Method	Mean % change (SD)	95% Confidence interval	t-test p-value
FreeSurfer	0.007 (0.452)	-0.233, 0.249	0.944
DiReCT	-0.181 (1.769)	-1.161, 0.798	0.697
MLI	-0.265 (0.794)	-0.688, 0.158	0.201
LCT	-0.074 (0.773)	-0.485, 0.338	0.708

Sensitivity analysis: We next examined the sensitivity of each method in detecting cortical changes globally and regionally where the AD group is expected to exhibit greater cortical changes than the normal. Table 2 summarizes the results performed on the global level. For all four methods, the average cortical thickness for the normal group was larger than that for the AD subjects, as expected. The change of -1.442% in the AD group as measured by our method is consistent with previously reported rates of 1 to 3% [7, 6]. For comparison, DiReCT measured the largest change (-3.299%), followed by our method (-1.442%), FreeSurfer (-0.811%) and MLI (-0.346%). Our method in particular measured a mean cortical thickness of 2.540 mm (SD = 0.136) for normals and 2.393 mm (SD = 0.088) for AD.

To measure the significance of the detected cortical thickness changes, we performed three tests of group separability (between AD and normal). Firstly, a series of student's t-tests show that *only* our method was able to differentiate the two groups with statistical significance. Secondly, we also performed a Wilcoxon rank-sum test, as the t-test is quite sensitive to the presence of outliers. Again, results show that *only* our method detected significant differences between the two groups. Thirdly, we also measured the effect size of each method using Cohen's d-test and found that LCT produced the highest effect size (0.441) while those of MLI, DiReCT

Table 2: Mean cortical thickness change (%) over 1 year computed by 4 methods in 50 AD and 50 normal subjects. The results of two-tailed t-test and Wilcoxon rank-sum test of group separability are also given. The p-values show that only our method was able to produce statistically different means between AD and healthy controls.

Method	Normal Mean % change (SD)	AD Mean % change (SD)	t-test p-value	Wilcoxon p-value
FreeSurfer	-0.418 (1.959)	-0.811 (2.741)	0.417	0.126
DiReCT	-1.761 (5.085)	-3.299 (5.273)	0.145	0.215
MLI	0.206 (1.989)	-0.346 (1.352)	0.105	0.237
LCT	-0.145 (2.556)	-1.442 (3.320)	**0.033**	**0.015**

Table 3: The p-values of the left and right hemispheres show that our method and DiReCT were able to measure statistically different means between AD and healthy controls in the frontal and temporal lobes whereas only the right temporal lobe was significant with FreeSurfer (the boldfaced p-values indicate $p < 0.05$).

Method	Frontal	Temporal	Parietal	Occipital
FreeSurfer	0.781, 0.652	0.151, **0.002**	0.769, 0.260	0.277, 0.191
DiReCT	0.051, **0.035**	**0.044, 0.013**	0.307, 0.081	0.625, 0.463
LCT	**0.012, 0.020**	**0.027, 0.048**	0.284, 0.293	0.732, 0.302

and FreeSurfer were 0.327, 0.206 and 0.165, resp.

Further, we performed a power analysis to evaluate the effectiveness of each method in detecting changes. Higher power means that the method can detect subtle changes using a smaller sample size with a given degree of confidence. From the analyses, we observed that the LCT demonstrated higher power (0.56) than MLI (0.35), DiReCT (0.30) and FreeSurfer (0.18). We also repeated the power analysis using bootstrapping (with replacement) to evaluate the ability of the measurement methods to produce separable means as the sample size is varied. The sample size was varied from 2 to 50, and 999 bootstrap iterations were performed for each sample size. The results (Figure 2) show that our method has the highest power for this data for virtually the whole range of sample sizes. This shows that our method is more sensitive than other state-of-the-art methods and for a fixed power, our method requires the smallest sample size to detect a significant change.

We also investigated the agreement between our method and the others by measuring Pearson correlations between them. Our method correlated strongly with MLI (r = 0.799, $p < 0.001$), moderately with DiReCT (0.573, $p < 0.001$), and modestly with FreeSurfer (0.231, $p < 0.05$). The good agreement of LCT with MLI and DiReCT is likely due to all three being voxel-based methods. The particularly high correlation between LCT and MLI is probably due to the use of the skeletal representation in both cases, with the increased sensitivity shown by our method being a likely consequence of the cortex-specific feature matching. Overall, the strong correlation of LCT with DiReCT suggests that our method gave results consistent with those of another voxel-based method.

We next performed regional analyses but omitted MLI this time since it is a cross-sectional method and it gave inferior performance on the global level. Table 3 reports the significance of the detected changes in terms of p-values, with our method and DiReCT showing overall greater sensitivity in the frontal and temporal lobes in both hemispheres than FreeSurfer, which was not sensitive to change except in the right temporal lobe. Note that our measurements are again consistent with earlier AD studies, where significant differences have been found in the frontal and temporal lobes [12, 13]. Finally, we also found strong correlations between LCT and DiReCT, with the frontal and temporal lobes showing the strongest correlations, ranging from 0.6 to 0.7.

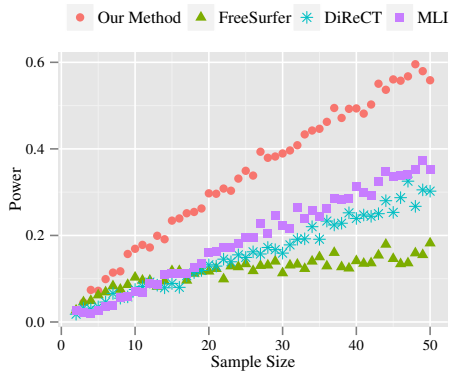

Fig. 2: Plot showing the change in power as the sample size is increased in bootstrapping for four different methods (with replacement). Our proposed method consistently shows higher power than other four methods for different sample sizes.

Computational efficiency: On average, our proposed method (LCT) took approximately 2.5 hours on an Intel Core2 Quad Q8200 2.33 GHz machine with 8 GB of RAM to process a pair of scans, while FreeSurfer and DiReCT took about 26 and 20 hours, respectively. Note that DiReCT was tested using two CPU threads on a computing cluster network, while our method and FreeSurfer were run only using a single thread. Regardless, our proposed method is approximately 10x faster than FreeSurfer and 8x faster than DiReCT.

4. DISCUSSION AND CONCLUSION

We have presented a novel method to perform longitudinal measurements of changes in cortical thickness. Unlike other methods, our proposed method does not require deformable registration, but uses a robust feature-matching approach specifically targeting the cortex. Fuzzy correspondence is used to enhance the stability of matches, with the results demonstrating that our method is more sensitive to the difference in atrophy rates between AD patients and normal controls than MLI. While the lack of ground truth makes resolving differences between different methods difficult, we have performed numerous experiments to evaluate LCT, and have found that our proposed approach has: 1) reproducibility comparable to FreeSurfer and MLI, and better than DiReCT; 2) sensitivity comparable to DiReCT, yet higher than FreeSurfer and MLI; and 3) higher computational efficiency than FreeSurfer and DiReCT, and comparable to MLI. Overall, our approach has demonstrated significant advantages for large longitudinal neurological studies.

5. ACKNOWLEDGEMENTS

This work was supported by grant funding from NSERC and the Milan and Maureen Ilich Foundation.

6. REFERENCES

[1] F. Liem, S. Mérillat, L. Bezzola, S. Hirsiger, M. Philipp, T. Madhyastha, and L. Jäncke, "Reliability and statistical power analysis of cortical and subcortical FreeSurfer metrics in a large sample of healthy elderly," *NeuroImage*, vol. 108, no. 0, pp. 95 – 109, 2015.

[2] K. Nakamura, R. Fox, and E. Fisher, "CLADA: cortical longitudinal atrophy detection algorithm." *NeuroImage*, vol. 54, no. 1, pp. 278–89, Jan 2011.

[3] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, "Within-subject template estimation for unbiased longitudinal image analysis," *NeuroImage*, vol. 61, no. 4, pp. 1402–1418, 2012.

[4] S. R. Das, B. B. Avants, M. Grossman, and J. C. Gee, "Registration based cortical thickness measurement." *NeuroImage*, vol. 45, no. 3, pp. 867–79, Apr 2009.

[5] Y. Li and D. Shen, "Consistent 4D cortical thickness measurement for longitudinal neuroimaging study," *MICCAI*, vol. 13, no. Pt 2, pp. 133–142, 2010.

[6] M. J. Clarkson, M. J. Cardoso, G. R. Ridgway, M. Modat, K. K. Leung, J. D. Rohrer, N. C. Fox, and S. Ourselin, "A comparison of voxel and surface based cortical thickness estimation methods." *NeuroImage*, vol. 57, no. 3, pp. 856–65, Aug 2011.

[7] I. Aganj, G. Sapiro, N. Parikshak, S. K. Madsen, and P. M. Thompson, "Measurement of cortical thickness from MRI by minimum line integrals on soft-classified tissue," *Human Brain Mapping*, vol. 30, no. 10, pp. 3188–3199, 2010.

[8] S. Garg, A. Traboulsee, D. K. B. Li, and R. Tam, "A robust and sensitive voxel-based method for measuring cortical thickness change using fuzzy correspondence," *In Proceedings of the XXIV Brazilian Congress of Biomedical Engineering*, pp. 1617–1620, 2014.

[9] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 24, pp. 509–522, 2002.

[10] M. Scott, P. Bromiley, N. Thacker, C. Hutchinson, and A. Jackson, "A fast, model-independent method for cerebral cortical thickness estimation using MRI," *Medical Image Analysis*, vol. 13, no. 2, pp. 269–85, Apr 2009.

[11] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The Alzheimer's Disease Neuroimaging Initiative," *Neuroimaging Clin N Am.*, vol. 15, no. 4, pp. 869–877, 2005.

[12] O. Acosta, P. Bourgeat, M. A. Zuluaga, J. Fripp, O. Salvado, S. Ourselin, and A. D. N. Initiative, "Automated voxel-based 3D cortical thickness measurement in a combined Lagrangian-Eulerian PDE approach using partial volume maps," *Medical Image Analysis*, vol. 13, no. 5, pp. 730–743, 2009.

[13] V. Singh, H. Chertkow, J. P. Lerch, A. C. Evans, A. E. Dorr, and N. J. Kabani, "Spatial patterns of cortical thinning in mild cognitive impairment and alzheimer's disease," *Brain*, vol. 129, no. 11, pp. 2885–2893, 2006.