Beyaz Cevher Hiperintensitesinin Görsel Derecelendirmesinde Otomasyon, Standardizasyon ve Tutarlılık Çağrısı
A Call for Automation, Standardization and Consistency in White Matter Hyperintensity Visual Grading

Leonardo O. Iheme[1], *Melek Kandemir*[2], *Z. Betul Yalciner*[2], *Devrim Unay*[3]

[1] Department of Electrical and Electronics Engineering, Bahcesehir University
leonardo.iheme@ieee.org, devrim.unay@ieu.edu.tr
[2] Department of Neurology, Bayındır Hastanesi İçerenköy
[3] Department of Electrical and Electronics Engineering, Izmir University of Economics

Özet

Uzman görsel değerlendirme yöntemlerinin uyumsuzluğunu ve öznelliğini gösteren analizler sunarak otomatik beyaz cevher hiperintensite derecelendirme sistemi ihtiyacını vurguluyoruz. Bu nihai amaca ulaşmak için otomatik bölütlemeye ihtiyaç vardır ancak uzmanların elle bölütlemeleri arasındaki yüksek uyum hatasız otomatik bölütlemenin ulaşılmaz olmadığına işaret etmektedir. Bunu yöntemimizle elde ettiğimiz çok yüksek Dice değerleri ile göstermekteyiz.

Abstract

We stress the need for an automated white matter hyperintensity rating system by showing analyses that expose the inconsistency and subjectivity of the expert visual rating method(s). To achieve the ultimate goal, automatic segmentation is required but the strong agreement between expert manual segmentations indicate that accurate automatic segmentation is not far fetched. We demonstrate this by the very good dice values achieved by our algorithm.

1. Introduction

White matter hyperintensity rating is important in monitoring and partly discretizing the progression of a number of diseases including Multiple Sclerosis (MS) and Alzheimer's Disease (AD) among others. The importance of rating seems to be on the increase and the development and analyses of several rating scales by many researchers from various groups around the world is an indicator. [1] compared 13 different rating scales; in [2] the degree of heterogeneity between six different scales was investigated and in [3] three rating scales were compared. Though studies have reported relatively high inter-scale agreement, they have also expressed concerns about the assumptions that were made while the conversion from one scale to another was carried out: for example, [2] applied a linear regression to partly non linear scales.

Bu çalışma Tübitak tarafından 111E083 nolu proje kapsamında, L.O.Iheme de BIDEB 2215 bursu ile desteklenmektedir.

The abundance of these scales could lead to inconsistent conclusions [1]; even more worrying is the fact that though raters are required to be well trained and experienced in order to be accurate and consistent while rating patients, it is not always the case. The issue of inter-rater and even intra-rater agreement arises and needs to be tackled. Thus we propose the use of an automated rating method that should be capable of consistently and accurately rating patients' white matter lesions.

An indispensable aspect of rating is segmentation. Even for visual ratings, experts perform a mental segmentation so as to isolate the regions of interest. For research among other purposes, experts have in many cases manually segmented lesions to produce masks that serve as ground truth for comparison. In this sphere, experts seem to reach high degrees of agreement across sites; [4] and [5] reported up to 0.83 and 0.98 agreement and ICC values respectively. Based on the fairly consistent and accurate manual delineations, the automation of segmentation has found good basis for comparison and evaluation.

This study further exposes the need for an automated rating system and contributes an expertly rated database. The rest of the paper is organized as follows: in section two, the materials and methodology used in the study are briefly described followed by the results, discussions and conclusion in sections three, four and five respectively.

2. Materials and Methods

In this section we will briefly describe the materials and methods employed in this study.

2.1. Data

The data used in this study was obtained from the ISBI MS Challenge 2015. It consists of the longitudinal MRI scans of five subjects manually segmented by two expert raters. Each time point contains T1-, T2-, and T2-weighted FLAIR MRIs. The segmentations were obtained from the pre-processed T2 FLAIR images that were also made available to challenge participants. For details of the pre-processing pipeline please refer to [6].

In addition to the provided data, our team of experts assessed and rated each image according to the in-house developed rating system that has been in use for research purposes at the polyclinic.

2.2. The Grading System

The scale on which patients are rated is developed from [7] but with slight modifications: the lower grade is further divided into two so that patients with little or no lesions are assigned a separate grade from those that have a *caps* grade. Exemplary images of each grade are shown in Figure 1 and a more detailed description of the grades can be obtained from [8].

2.3. Automatic Segmentation of White Matter Hyperintensity

2.3.1. Intensity Thresholding

We begin by mapping the intensities of every training image to those of a reference image which in this case is the first image of subject 1; we then compute the histogram of the whole brain foreground voxels from the FLAIR image and assume that the peak is that of a normal distribution so that its 7 dB drop is more than twice its Full Width at Half Maximum (FWHM). The intensity of this I_{7dB} point is guaranteed to be amongst the highest intensity values of the image. With this value as a minimum threshold for WMH, we define the threshold as:

$$T = I_{peak}(1 - w) + I_{7dB} \quad (1)$$

where w is a weight value to be determined. For a more detailed description and an evaluation of the method, refer to [8].

2.3.2. 3D Connectivity Analysis & Corpus Callosum Delineation

The 3D connectivity analysis in brief involves the examination of every detected voxel for the degree of connectivity with its neighboring voxels. This step translates to analyzing the volumetric significance of every detected lesion. Lesions that are deemed insignificant are assumed to be false positives and are therefore ignored during the segmentation. A parameter that needs to be optimized during training is the minimum volume of lesions. Refer to Figure 2 where the false positive minimization effect of the 3D connectivity analysis is depicted.

Additional steps were required to get rid of false positives that appear along the corpus callosum. We employ a RANSAC based approach [9] to detect the inter-hemispheric fissure (IHF) which coincides with the corpus callosum when the image is viewed axially. Falsely detected voxels along the corpus callosum are then removed by excluding voxels that fall on the inter-hemispheric fissure line from the segmentation. Figure 3 briefly demonstrates the working of the RANSAC based IHF detection.

3. Results

It is well established that regarding manual segmentation, experts agree to a great degree and this is further demonstrated by the excellent agreement between the manual segmentations

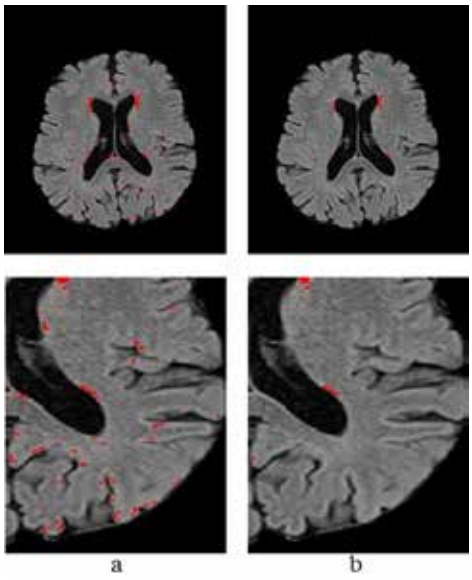

Figure 2: The effect of the 3D connectivity analysis. Figures a & b depict the before and after effects respectively

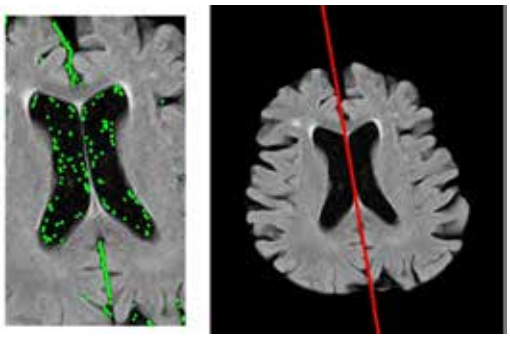

Figure 3: Inter-hemispheric fissure detection. The green dots on the right image are the candidate points to be used by the RANSAC algorithm for fitting the line of best fit that coincides with the corpus callosum

of the two experts. A mean dice score of 0.84 with a standard deviation of 0.06 was observed. Figure 4 depicts a sample manual segemntation performed by two experts. Regarding visual rating however, experts seem to disagree to a large extent as depicted in Figure 5 where the observed agreement between raters was 48% and the Kappa value was 0.2.

3.1. Expert Rater Analyses

Our experiments were setup as follows: the data was rated by two very experienced raters and their ratings were compared and reported as observed agreement. The subjects and their respective grades as assigned by the raters are presented in Table 1.

The segmentation algorithm introduced earlier in the paper yielded a mean dice score of 0.56 and a standard deviation of 0.21. At best, the algorithm was able to reach a dice score of 0.85 and at worst, the dice score was 0.11. Furthermore, Figure

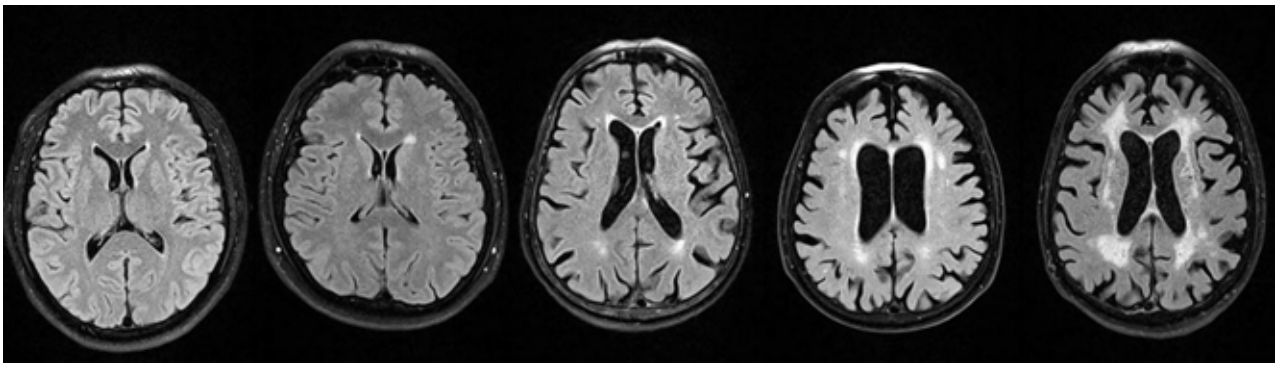

Figure 1: Sample images of individual grades. From left to right: Grade 0, Grade 1 (*caps*), Grade 2 (*thin line*), Grade 3 (*halo*), Grade 4 (*diffused*)

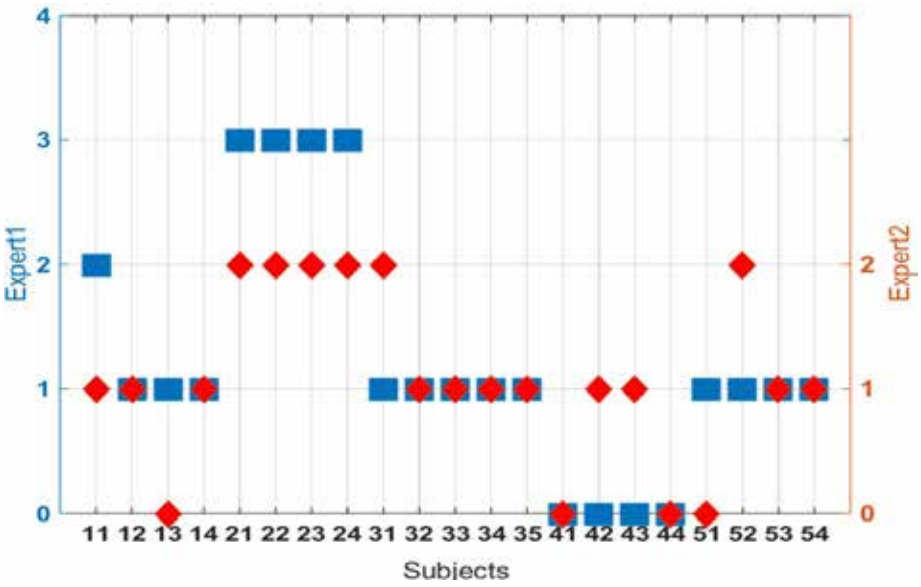

Figure 5: Expert raters grade distribution. The blue square markers are those of *Expert1* while the red diamond markers are those of *Expert2*

2 depicts a sample visual result of how the algorithm performs on a single slice of one of the subjects.

4. Discussion

The results presented reveal a strong agreement between experts' segmentations however, when the data was rated, a poor agreement was observed. The revelations further buttress the age old fact that visual expert rating of WMH is highly subjective and could be inconsistent; they however strengthen the need for there to be a standardized and consistent rating system. Although our algorithm performed badly on subjects with lower grades, we can easily infer that the goal of achieving a reliable automated segmentation method is etching closer as demonstrated by the high dice score achieved. A more robust algorithm is required in order to successfully segment WMHs of subjects with small lesions. It is important to point out that such a task has proven to be quite difficult as demonstrated in [6] where the algorithms presented also performed relatively poorly on the same subject when compared to the results of the other subjects.

It is essential to observe from Table 1 that the expert raters' assigned grades never differed by more than one grade in fact the two raters are fairly correlated (0.68) though Expert1 seemed to assign lower grades than Expert2 in general. This observation exposes the need for a standardization of the rating system but due to the fact that human observation and judgment is subjective, the discrepancy may not be resolvable if we continually rely solely on visual grading of patients.

5. Conclusion

In this work we have briefly introduced the grading system developed from [7], used at our local polyclinic and applied it to a dataset provided by [6]. The agreements between the ratings which were performed by two experienced experts in the field exposed the subjectivity and inconsistency of the rating system in general. Thus we proposed the automation of the process and

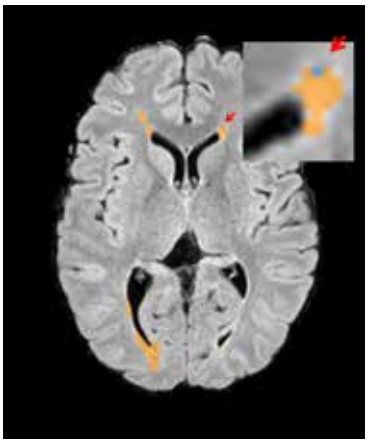

Figure 4: Manual segmentation of two experts. The orange overlays represent the regions where the two segmentations overlap while the blue regions (indicated by the red arrowhead) show mismatch

went a step further in automatically segmenting lesions from the same dataset. Our segmentation algorithm performed very well on subjects with large lesions but exposed the need for more robust algorithms especially for subjects with tiny lesions.

While rating WMH is highly advantageous in the assessment of the level of severity of the disease in a patient, the need for a high level of consistency cannot be overemphasized. Perhaps a machine learning based approach to automatic rating would be worth investing into.

6. References

[1] R Mäntylä, T Erkinjuntti, O Salonen, H J Aronen, T Peltonen, T Pohjasvaara, and C G Standertskjöld-Nordenstam. Variable agreement between visual rating scales for white matter hyperintensities on MRI. Comparison of 13 rating scales in a poststroke cohort. *Stroke.*, 28(8):1614–1623, August 1997.

[2] Leonardo Pantoni, Michela Simoni, Giovanni Pracucci, Reinhold Schmidt, Frederik Barkhof, and Domenico Inzitari. Visual rating scales for age-related white matter changes (leukoaraiosis): Can the heterogeneity be reduced? *Stroke*, 33(12):2827–2833, 2002.

[3] P. Kapeller, R. Barber, R. J. Vermeulen, H. Adèr, P. Scheltens, W. Freidl, O. Almkvist, M. Moretti, T. Del Ser, P. Vaghfeldt, C. Enzinger, F. Barkhof, D. Inzitari, T. Erkinjunti, R. Schmidt, and Franz Fazekas. Visual rating of age-related white matter changes on magnetic resonance imaging: Scale comparison, interrater agreement, and correlations with quantitative measurements. *Stroke*, 34(2):441–445, January 2003.

[4] D. M J Van Den Heuvel, V. H. Ten Dam, A. J M De Craen, F. Admiraal-Behloul, A. C G M Van Es, W. M. Palm, A. Spilt, E. L E M Bollen, G. J. Blauw, L. Launer, R. G J Westendorp, and M. A. Van Buchem. Measuring longitudinal white matter changes: Comparison of a visual rating scale with a volumetric measurement. *Am. J. Neuroradiol.*, 27(4):875–878, 2006.

[5] F. Admiraal-Behloul, D. M J Van Den Heuvel, H. Olofsen, M. J P Van Osch, J. Van Der Grond, M. a. Van Buchem, and J. H C Reiber. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *Neuroimage*, 28(3):607–617, November 2005.

[6] Pham Dzung, Bazin Pierre-Louis, Carass Aaron, Calabresi Peter, Crainiceanu Ciprian, Ellingsen Lotta, He Qing, Prince Jerry, Reich Daniel, and Roy Snehashis. THE 2015 LONGITUDINAL MULTIPLE SCLEROSIS LESION SEGMENTATION CHALLENGE, 2015.

[7] Franz Fazekas, John B Chawluk, A Alavi, HI Hurtig, and RA Zimmerman. Mr signal abnormalities at 1.5 t in alzheimer's dementia and normal aging. *American Journal of Roentgenology*, 149(2):351–356, 1987.

[8] L.O. Iheme, D. Unay, O. Baskaya, A. Sennaz, M. Kandemir, Z.B. Yalciner, M.S. Tepe, T. Kahraman, and G. Unal. Concordance between computer-based neuroimaging findings and expert assessments in dementia grading. In *Signal Processing and Communications Applications Conference (SIU), 2013 21st*, pages 1–4, April 2013.

[9] Ahmet Ekin. Feature-based brain mid-sagittal plane detection by RANSAC, 2006.

Table 1: Subjects and assigned grades

Subject	Time point	Expert2	Expert1
1	11	2	1
	12	1	1
	13	1	0
	14	1	1
2	21	3	2
	22	3	2
	23	3	2
	24	3	2
3	31	1	2
	32	1	1
	33	1	1
	34	1	1
	35	1	1
4	41	0	0
	42	0	1
	43	0	1
	44	0	0
5	51	1	0
	52	1	2
	53	1	1
	54	1	1