3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation

Release 1.00

Martin Styner[1,2], Joohwi Lee[1], Brian Chin[3], Matthew S Chin[2], Olivier Commowick[4], Hoai-Huong Tran[4], Valerie Jewells[3], Simon Warfield[4]

August 25, 2008

[1]Dept. Computer Science, Univ. of North Carolina, Chapel Hill NC, USA
[2]Dept. Psychiatry, Univ. of North Carolina, Chapel Hill NC, USA
[3]Dept. of Neuroradiology, Univ, of North Carolina, Chapel Hill, NC, USA
[4]Dep. of Radiology, Children's Hospital Boston, Boston, MA

Abstract

This paper describes the setup of a segmentation competition for the automatic extraction of Multiple Sclerosis (MS) lesions from brain Magnetic Resonance Imaging (MRI) data. This competition is one of three competitions that make up a comparison workshop at the 2008 Medical Image Computing and Computer Assisted Intervention (MICCAI) conference and was modeled after the successful comparison workshop on liver and caudate segmentation at the 2007 MICCAI conference. In this paper, the rationale for organizing the competition is discussed, the training and test data sets for both segmentation tasks are described and the scoring system used to evaluate the segmentation is presented.

Contents

1 Introduction

This paper represents the introductory paper for the MICCAI 2008 workshop on 3D Segmentation in the Clinic focusing on Multiple Sclerosis (MS) lesion segmentation. MS is a demyelinating disease of the central nervous system and usually shows inammatory white matter (WM) pathology leading to various cognitive impairements. Magnetic Resonance Imaging (MRI) is one of the diagnostic imaging methods for MS. MRI's of MS patients primarily exhibit focal and less often diffuse WM lesions in the brain and spinal cord. Using computing tools, the lesion load is often analyzed and longitudinally monitored for patient follow-up and drug efficacy studies. As for many other segmentation tasks, manual delineation of MS lesions is the still the routine, but is also both challenging, variable and time-consuming.

A wide variety of methods are available for automatic and semi-automatic MS lesion segmentation. Currently methods modeling the lesions either directly or as outliers in tissue classification seem most widely used, but also a variety of methods that do not fit into this categories have been proposed and are in use. Not only is the number of available methods increasing, more and more methods are, or claim to be, generic and applicable to multiple segmentation tasks, usually after applying some suitable modifications or tweaks. Even for experienced researchers in the field it is difficult to choose the appropriate technique for a particular problem. WIth this workshop we aim at generating information for comparing MS lesion segmentation methods in an unbiased and honest fashion.

The setup of this competition follows the first installment of this MICCAI workshop [2]. That workshop had two parts focusing on liver and caudate segmentation, where as this year's workshop has three parts focusing on coronary artery central lumen line extraction, this MS lesion segmentation, and liver tumor segmentation.

The competition serves purposes other than as a comparative study of a range of algorithms on a common database. It also provides a snapshot of currently popular methods for medical image analysis. Moreover, for many researchers it is difficult to obtain a sizeable amount of training and test scans with high quality segmentations. From their feedback, we concluded that in many cases the data sets of the workshop allowed them to test their work on new data, which sparked ideas to improve their algorithms.

2 General Organization

Any team could enter the competition after signing a letter of intent. Only fully automatic methods were allowed to participate. For this competition three sets of data were constructed. The first is a training dataset that includes both images and binary masks of the segmentations of the MS lesions, produced by human experts. Participants can use these images to train their algorithm, but they are free to use their own training data in addition to the supplied data. They may also use the training data to determine suitable values for free parameters in their algorithms. The second data set is a test set that was distributed with the training data (both could be downloaded from a web site). The test set did not include segmentations and participants had to send in the segmented test sets before a given deadline.

Finally there is a second test set that will be released at the start of the one-day workshop on September 6th, 2008. This set is the onsite test set. Segmentation results are to be submitted within three hours after the onsite test set has been made available. The actual onsite test data was made available before the workshop as a password protected zip file and the password will be distributed at the workshop. Teams containing members of the workshop organizers were excluded from this onsite competition.

The reason for using two different test sets is that distributing the test data to be segmented together with the training data allows teams to optimize their algorithms for the provided test cases. This can be the case even

if the segmentations of the test cases are unknown, and can happen even unintentionally. A contest during the actual meeting on new test data also increased the competitive element of the workshop.

All submitted segmentations were processed with the same fixed evaluation protocol described further below.

3   Training and Testing Database

The databases consisted of scans from research subjects acquired at UNC (3T Siemens Allegra) and CHB (3T Siemens). All datasets were fully anonymized for dissemination purposes. While eventually all datasets were segmented by a single expert rater at CHB (HT) and jointly by 2 expert raters at UNC (BC, MC), only the CHB segmentation data was available to the workshop participants on-time for training their algorithms. The testing evaluation was performed against both the UNC and CHB expert results.

The following distribution of the data was provided to the participants:

- 20 training cases: 10 CHB and 10 UNC cases were provided with manual segmentations from the CHB expert.

- 25 testing cases: 15 CHB and 10 UNC cases were provided without expert segmentations.

- The datasets were randomly assigned to training and testing, while maintaining site origin.

For all cases, the database contained the same number of high resolution images: A T1 weighted scan, a T2 weighted image, a FLAIR image, a Diffusion Tensor Imaging (DTI) derived Fractional Anisotropy (FA) and Mean Diffusivity (MD) image. Prior to dissemination, all data sets were re-oriented to axial orientation. The T1 image was then rigidly registered to the standard MNI atlas. The T2 weighted and FLAIR images were rigidly registered to its corresponding T1 images. The DTI data was non-rigidly registered to the corresponding T2 images via b-spline based normalized mutual information registration[1]. All images were resliced at isotropic 0.5x0.5x0.5mm resolution with cubic spline interpolation. This resolution was chosen, as most of the structural T1, T2 and FLAIR datasets had originally an in-plane resolution of 0.5x0.5 and several images had further an original slice thickness of 0.5mm. This means that many scans though were up-interpolated to a higher resolution than the original one. Datasets were randomized within its site origin and workshop participants were blind to the randomization. As dataformat we chose the ITK-readble NRRD format that is comprised of an ASCII readable header file and the separate uncompressed raw image data file as this allows straightforward reading with ITK but also easy readability if participants chose other I/O routines.

4   Evaluation

The quality of a segmentation can be evaluated in many different ways. A sensible evaluation criterion depends on the purpose of the segmentation procedure. If the goal is to estimate the volume of MS lesions, a measure often referred to as lesion load, the volumetric error would be an obvious criterion. But a segmentation with exactly the same volume as the reference (we refer to the expert manual segmentations that we consider to be the reference) can be completely wrong if a voxel by voxel comparison is made. As a result, there are different evaluation metrics in common use and most papers on segmentation report results in terms of more than one metric.

As this workshop represents a competition, we decided to compute a single evaluation criterion that incorporates several popular metrics. To be able to give a meaningful interpretation to these scores, we decided to gauge each metric by relating it to the result that could be expected if an independent human observer would perform the segmentation manually. We also wanted to avoid that a few completely failed segmentations would damage an overall score irreparably. Thus we decided to award 100 points for a perfect result (the best value that could be obtained for a metric) and a predefined amount of 90 for a score that is typical for an independent human observer. The scaling between these two gauge values is linear, but negative scores are impossible, so 0 points is the minimum possible value to achieve.

The following four error metrics were used:

- Relative absolute volume difference, in percent: The total absolute volume difference of the segmentation to the reference is divided by the total volume of the reference, in percent. Note that the perfect value of 0 can also be obtained for a non-perfect segmentation, as long as the volume of that segmentation is equal to the volume of the reference.

- Average symmetric surface distance, in millimeters: The border voxels of segmentation and reference are determined. These are defined as those voxels in the object that have at least one neighbor (of their 18 nearest neighbors) that does not belong to the object. For each voxel along one border, the closest voxel along the other border is determined (using unsigned Euclidean distance in real world distances, thus taking into account the different resolutions in the different scan directions). All these distances are stored, for border voxels from both reference and segmentation. The average of all these distances gives the averages symmetric absolute surface distance. This value is 0 for a perfect segmentation.

- True Positive Rate, in percent: This is measured by dividing the number of lesions in the segmentation that overlap with a lesion in the reference segmentation with the number of overall lesions in the reference segmentation. This evaluates whether all lesions have been detected that are also in the reference segmentation. It is though possible to have a perfect score of 100% and have additional lesions as compared to the reference segmentation. A caveat for this measurement is further that if correctly detected lesions are fused as compared to the reference segmentation, then this is registered as a partial error.

- False Positive Rate, in percent: This is measured by dividing the number of lesions in the segmentation that do *not* overlap with any lesion in the reference segmentation with the number of overall lesions in the segmentation. This rate represents whether any lesions are detected that are not in the reference. A method that oversegments lesions would have a low value for this method, whereas a very conservative method would have high values. An empty image would always score a perfect score of 0%.

Using this scoring system one can loosely say that 90 points for a MS lesion segmentation is 'comparable to a human expert performance'. But this is only a rough indication, note that the scores of humans will vary across cases, and across humans. For difficult cases, a score below 90 points may therefore still be very good. We have not investigated this in detail for the data sets used in this competition. To gauge the scoring system, three expert raters (1 CHB and 2 UNC based expert raters) segmented all of the UNC training cases. The UNC experts adapted their segmentation protocol to match closely to the one employed at CHB. Thus, while 2 sites are involved here, the measured expert variability comes close to a single protocol based inter-rater variability.

As the whole testing database was segmented by a single expert rater at CHB (HT) as well as jointly by 2 expert raters at UNC (BC, MC), 2 full sets of expert segmentations were used as reference for the comparisons. We expected methods to do better against the CHB rater, as the training provided to the participants

only consisted of segmentation from that same CHB rater. Generally, this did not happen, which suggests that the raters provided consistent, high quality segmentations.

In addition to the 2 rater segmentations, we computed a composite segmentation via the well-know STAPLE algorithm [3]. The input for STAPLE was all the rater segmentations, as well as the segmentations provided by the workshop participants. It thus represents a composite of two human experts (= the variable ground truth) and nine automatic segmentation methods. The resulting STAPLE segmentation is not used in scoring the methods, but serves as an additional evaluation. The following STAPLE measurements are reported alongside the scores:

- Sensitivity: This measurement represents a method's ability to segment the correct lesions and is correlated with the true positive rate as compared to the STAPLE composite segmentation.

- Specificity: This measurement represents a method's ability to avoid incorrect lesions and is correlated with the false positive rate as compared to the STAPLE composite segmentation.

- Positive predictive value (PPV). This measurement is the ratio of true positives to the sum of true positives and false positives, and represents a good compromise combining both sensitivity and specificity.

The scores and STAPLE measures were computed by the workshop organizers after participating teams had uploaded results to the workshop website http://www.ia.unc.edu/MSseg. Each team received a table with the scores to be included in their respective papers submitted to the proceedings. The evaluation software used in this workshop was also available online as open source as well as binaries for Linux, Windows and MacOSX.

5   Discussion

Despite the short period between the data availability and the deadline for the submission of the results, an unexpected large number of 30 teams registered and downloaded the data for the MS lesion segmentation workshop. About a third of the teams submitted results, many within a similar score range, coming close to inter-rater variability. We believe this response shows that there is a definite interest within the medical image segmentation research community to participate in comparative studies such as these. We intend to make the data sets and the results of the various systems described in these proceedings publicly available after the workshop on its homepage http://www.ia.unc.edu/MSseg.

6   Acknowledgment

References

[1] D Rueckert, L I Sonoda, C Hayes, D L Hill, M O Leach, and D J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Transactions On Medical Imaging*, 18(8):712–21, Aug 1999. 3

[2] Bram van Ginneken, Tobias Heimann, and Martin A Styner. 3d segmentation in the clinic: A grand challenge. *Workshop on 3D Segmentation in the Clinic, MICCAI 2007*, pages 7–15, 2007. 1

[3] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–21, Jul 2004. 4