An Algorithm for Automatic Segmentation and Classification of Magnetic Resonance Brain Images

Bradley J. Erickson and Ramesh T.V. Avula

In this article, we describe the development and validation of an automatic algorithm to segment brain from extracranial tissues, and to classify intracranial tissues as cerebrospinal fluid (CSF), gray matter (GM), white matter (WM) or pathology. T1 weighted spin echo, dual echo fast spin echo (T2 weighted and proton density (PD) weighted images) and fast Fluid Attenuated Inversion Recovery (FLAIR) magnetic resonance (MR) images were acquired in 100 normal patients and 9 multiple sclerosis (MS) patients. One of the normal studies had synthesized MS-like lesions superimposed. This allowed precise measurement of the accuracy of the classification. The 9 MS patients were imaged twice in one week. The algorithm was applied to these data sets to measure reproducibility. The accuracy was measured based on the synthetic lesion images, where the true voxel class was known. Ninety-six percent of normal intradural tissue voxels (GM, WM, and CSF) were labeled correctly, and 94% of pathological tissues were labeled correctly. A low coefficient of variation (COV) was found (mean, 4.1%) for measurement of brain tissues and pathology when comparing MRI scans on the 9 patients. A totally automatic segmentation algorithm has been described which accurately and reproducibly segments and classifies intradural tissues based on both synthetic and actual images.
Copyright © 1998 by W.B. Saunders Company

KEY TERMS: automatic multiparametric classification, brain segmentation, multiple sclerosis (MS), magnetic resonance imaging (MRI)

SEGMENTATION is the separation of the voxels of an image into groups. Classification identifies the nature or tissue type of each group. Segmentation and classification of magnetic resonance (MR) head images can provide a quantitative basis for evaluation of a disease process such as multiple sclerosis (MS).[1] Automated and semiautomated methods have considerable advantage over manual methods because of their objectivity and the time savings they allow.[2] Semiautomated techniques have been shown to be reasonably accurate and reproducible. However, they require user inter-

action early in the segmentation process, with some delay before the segmentation is completed. If segmentation and classification are to become a routine part of practice, it would be optimal to develop a totally automated process which can begin as soon as images are reconstructed and be completed before interpretation and reporting begin. In this article, we present such a totally automated process, apply it to both synthetic and real patient data, and demonstrate its accuracy and reproducibility.

METHODS

Image Acquisition and Preprocessing

The images for this study were 3 mm contiguous images obtained with 1.5T GE Signa imager (General Electric Medical Systems, Milwaukee, WI). The image sets consisted of spin echo T1 (450/17 TR/TE), dual echo fast spin echo PD and T2 (4000/17/102), and fast FLAIR (11000/139 TI 2600) acquisitions (Table 1). The mean and standard deviation (SD) of normal GM, WM, and CSF, for T1, PD, T2, and FLAIR images were obtained from 10 manual measurements in 10 normal patients.

For validating the accuracy of segmentation, a normal head MRI study was manually segmented into GM, WM, CSF, and extradural tissues. "Lesions" were created in this data set by drawing regions of interest on the normal human MR images. A Gaussian random number generator combined with measurements of actual voxel values of lesions for each of the image types was then used to replace the normal brain voxel values with lesion values. Whereas there is a wide variety in the appearance of MS lesions, the validity of the synthetic lesions was verified visually and numerically. This synthetic lesion image provides a very good gold standard because it is known precisely which voxels are pathological and which are normal.

To validate reproducibility, two complete MRI studies were obtained approximately 7 days apart in 9 chronic progressive MS patients. There was no change in the patient's clinical status during this time interval, nor had there been for the previous 6 months. Therefore, any change in lesion volume should be caused by variations owing to differences in patient position or the algorithm.

Images were registered by using a separate algorithm (similar to that of Ostuni et al[3]) that defined skin/subcutaneous fat interfaces on T1, PD and FLAIR images. The spatial transform that minimized the RMS error between the location of the surfaces was used to register images to within 1 voxel (verified by visual observation). The FLAIR image was used as the template for registration to minimize blurring it because it is used in the final determination of previously unclassified voxels. The T2 image was transformed by using the PD transformation matrix because they were acquired at the same time.

A multispectral anisotropic diffusion filter was applied to all MRI images to reduce noise without blurring edges.[4]

Table 1. Parameters Used for Each of the Four Image Types Used in This Study

	TR	TE (TI)
T1	450	17
Spin Density (SD)	4000	17
T2	4000	102
FLAIR	11000	139 (2600)

The Segmentation Algorithm

The key to the voxel classifier is the multiparametric space map. If image intensity values are normalized to the intensity of subcutaneous fat on each pulse sequence, the intensity of each normal intradural tissue type becomes essentially constant for each type of image. Mean voxel intensity values on 10 normal patient examinations were measured 10 times for subcutaneous fat, CSF, GM and WM on T1, T2, PD and FLAIR images. Parametric maps were generated by normalizing the images based on the intensity of subcutaneous fat. Figure 1 shows the ratio of the intensity of the tissues of interest versus fat for 9 normal patients. When fat intensity is normalized, the intensity of CSF, GM and WM are fairly fixed.

Subcutaneous fat is identified on T1 weighted images by finding large groups of contiguous peripherally located voxels that are above a threshold equal to 75% of the maximum voxel intensity on each slice. The corresponding voxels in the other sequences are labeled as fat because the sequences were spatially registered in a previous step. The mean intensity value for fat is calculated for each slice of each sequence. The voxels of each slice for each sequence are then scaled by the factor (1000/mean fat intensity). This slice-by-slice scaling corrects for field heterogeneity. Next, voxels are classified as CSF, GM, WM, or "other" by using the multiparametric spacemap. Three-dimensional (3D) erosion is applied to brain (GM + WM) voxels to remove bridges between brain and extradural tissues followed by 3D region growing of the largest collection of brain voxels (eliminating extracranial tissues with signal like brain). Conditional dilation is then applied to restore superficial brain voxels.

The final step is to classify all remaining unclassified intradural voxels. These could be unclassified if the underlying tissue is pathological or because of artifact (eg, misregistration, partial volume effects). In the first pass, the mean and SD of brain voxels (GM and WM) on FLAIR images is calculated. Any unlabeled voxels greater than 2 SDs above the mean are labeled as pathology, all those more than 2 SDs below the mean are labeled as CSF, and those remaining are labeled as brain. Because the FLAIR is used as the template for registration with the other pulse sequences, any minor misregistration of images or blurring of images during registration that inevitably occurs will not prevent labeling of voxels. The entire algorithm is described in pseudocode in Appendix A.
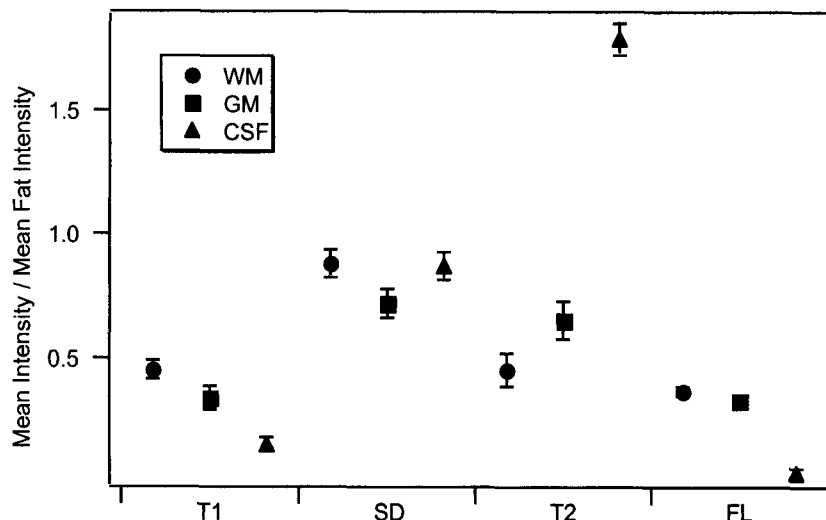
RESULTS

Normal and Synthetic Lesion Images

The results of the multiparametric classification algorithm for the manually segmented normal head MRI are shown in Figs 2 and 3. Ninety-four percent of brain (WM + GM) voxels were labeled correctly and 99% of CSF. Figure 3 shows a slice-by-slice comparison of the volume of brain and CSF for the gold standard based on manual segmentation. For the entire volume, 92% of WM, 96% of GM voxels, 99% of CSF, and 94% of pathology were labeled correctly on a voxel-by-voxel basis.

MS Patient Results

The initial and follow-up volumetric measurements of CSF, GM, WM and pathology in the 9 patients with MS are shown in Fig 4. Figure 5 shows a slice-by-slice comparison for one of the patients. The initial and follow-up examinations for this patient were reasonably well registered, and therefore, a slice-by-slice comparison is meaningful. The mean coefficient of variation (COV) for all

Fig 1. The ratio of tissue mean intensity to subcutaneous fat for white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) for each of the 4 pulse sequences used (T1, SD = spin density, T2, and FL = FLAIR). The error bars show plus and minus 2 standard deviations. The scanning parameters used (see Table 1) may affect these values.
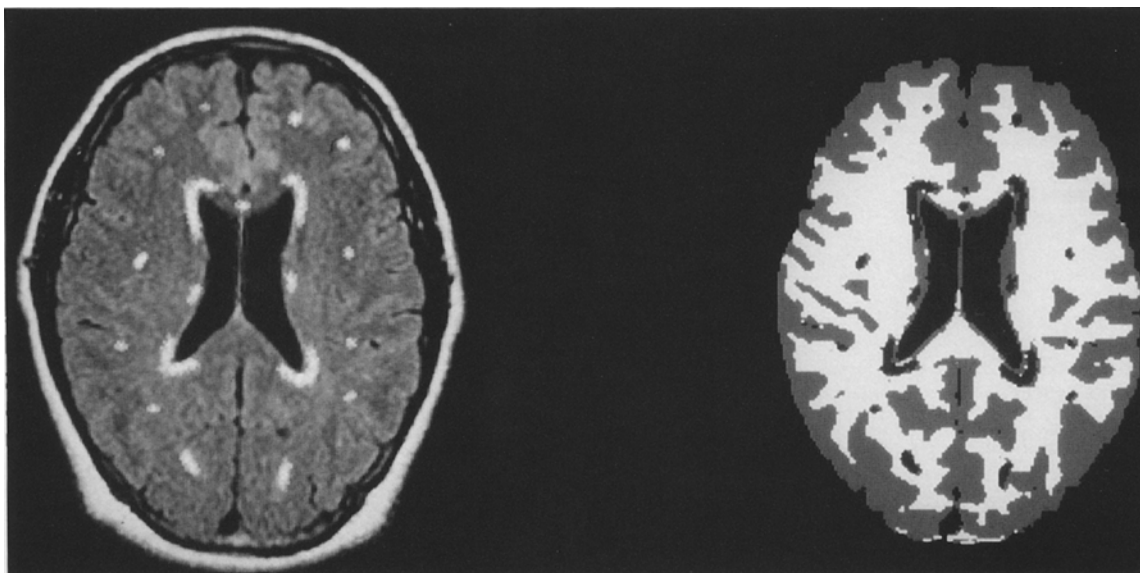
Fig 2.   Left image: FLAIR image with synthetic lesions. Right image: Segmented image.

studies was 4.1% and was always less than 12%. Figure 6 shows a graph of the COV for each of the 9 patients.

Performance

The algorithm (excluding registration) requires approximately 5 minutes to segment and classify a data set consisting of 54 3mm contiguous slices for each for each of the four images types (T1, PD, T2, and FLAIR), when executed on a 200 Mhz Pentium

Pro CPU (Intel Corp, Palo Alto, CA) with 128 MB RAM. The time required for the registration process is variable depending on the amount of patient motion between sequences. Typical times are 12 to 15 minutes, but may range from 5 to 50 minutes.

DISCUSSION

We have described a method for totally automatic segmentation of head MRIs. Although not providing a perfect result, we view it is an impor-
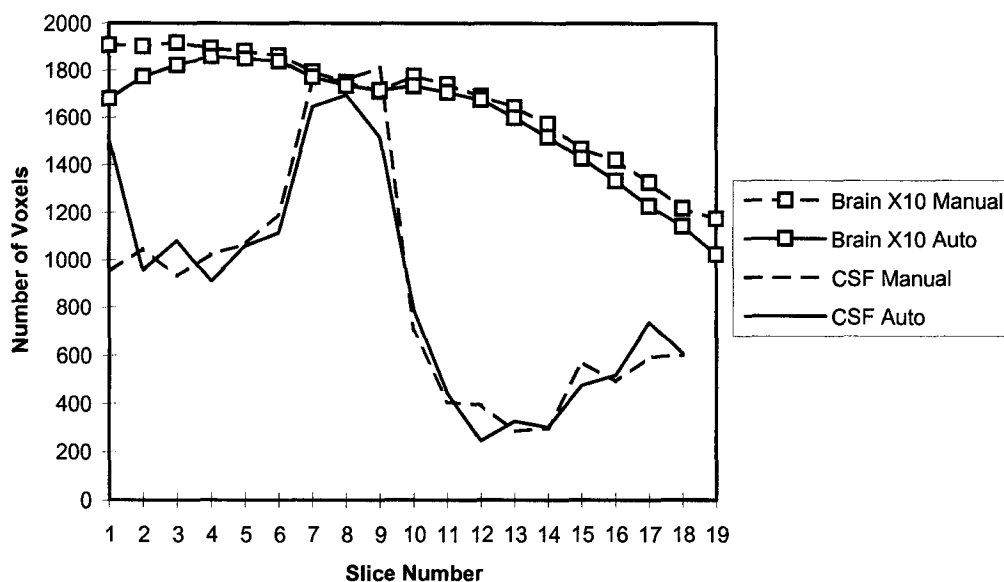

Fig 3.   .The results of the automated segmentation algorithm on the synthetic lesion images. The graph shows a slice-by-slice comparison of brain and CSF volumes determined manually and by the automatic algorithm.
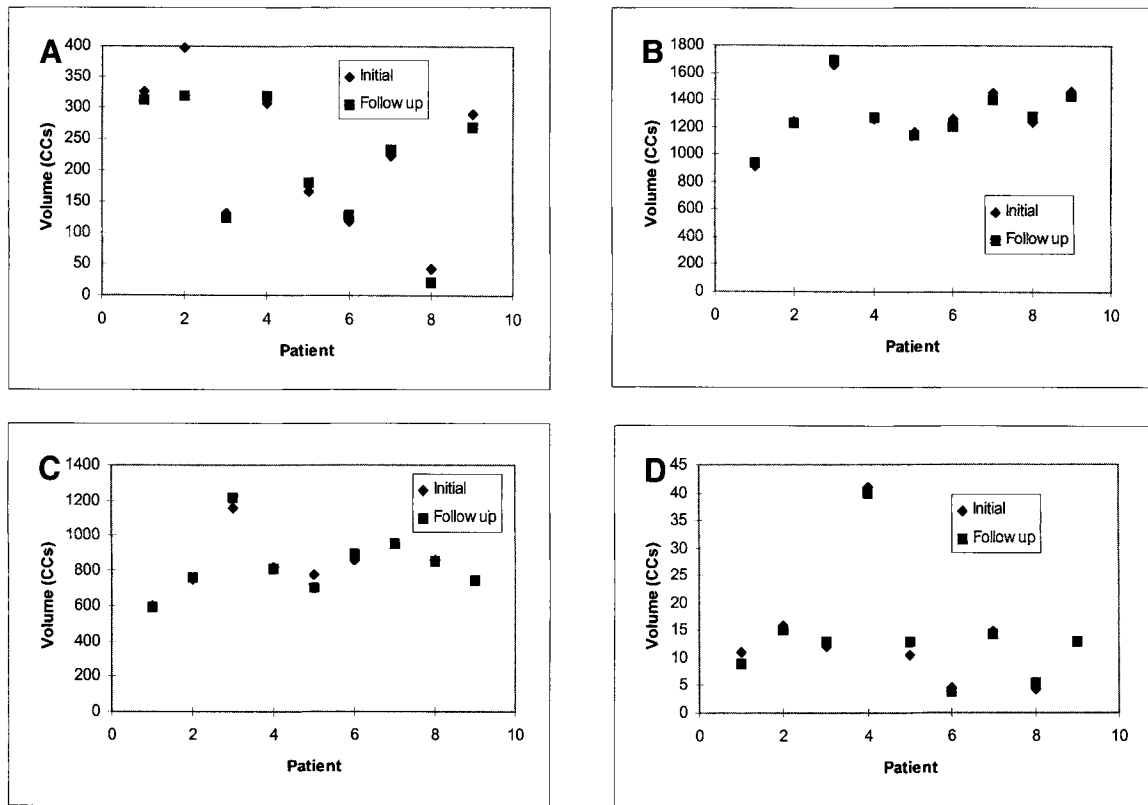
Fig 4. (A) Total volume of CSF, (B) WM, (C) GM, and (D) pathology for MS patients scanned at an interval of 1 week are compared.

tant step toward more timely and more widespread delivery of volumetric information into routine clinical practice. Clinical acceptance of any method will require accurate results and minimal operator intervention. Several investigators have proposed semiautomated segmentation schemes based on supervised and unsupervised techniques. These techniques require operator interaction at some stage. After that interaction, some processing is performed, and eventually the result is obtained. The time required to obtain measurements has hindered its application to routine practice. On the other hand, an automatic process can execute on any computer (including a remote server) as soon as the images have been reconstructed. This computer can then send the image labels to a radiology display station, making routine clinical use more feasible. It is quite possible to perform registration, classification, and segmentation on a study before most practices would be ready to interpret the study (less than 30 minutes after study completion).

Previous efforts have used semiautomatic segmentation and classification methods based on multiparametric cluster plot analysis,[5,6] as well as

nonparametric feature map techniques.[6] Most semiautomatic segmentation techniques can be categorized as supervised or unsupervised but both require operator interaction at some stage of segmentation. The supervised methods require the operator to select samples of each tissue class. Supervised segmentation techniques are commonly based on pattern recognition techniques such as the maximum likelihood method, k-nearest neighbors (k-NN), and back propagation artificial neural net algorithms. In a comparison of these methods in controls and patients with gliomas, k-NN provided best results.[7] To minimize user-induced variability, operator interaction should be minimized. In a recent study, unsupervised segmentation was preferable for measurement of tumor volume in response to treatment.[8] In that study, it was shown that the split fuzzy c-means algorithm identifies tissue better than the fuzzy c-means algorithm, which is mostly an unsupervised technique. Unfortunately, unsupervised techniques previously described still have the problem of identifying anatomically relevant tissue and determining how many classes there should be.
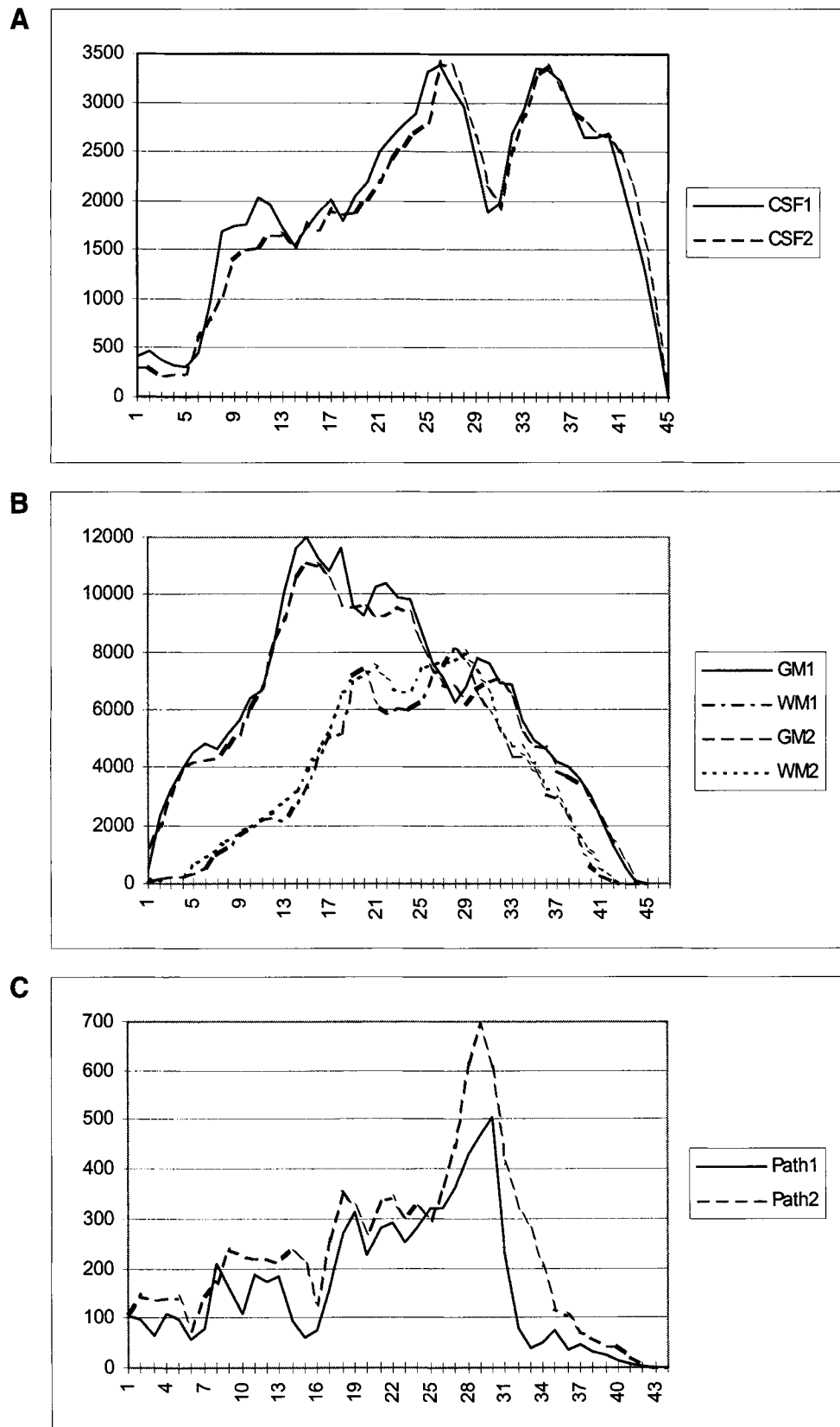
A

B

C

Fig 5. The number of voxels labeled as (A) CSF, (B) GM and WM, and (C) pathology were plotted for each slice for a patient scanned in interval of 1 week which were well registered. Lines marked 1 indicate initial scan and 2 indicate second scan.
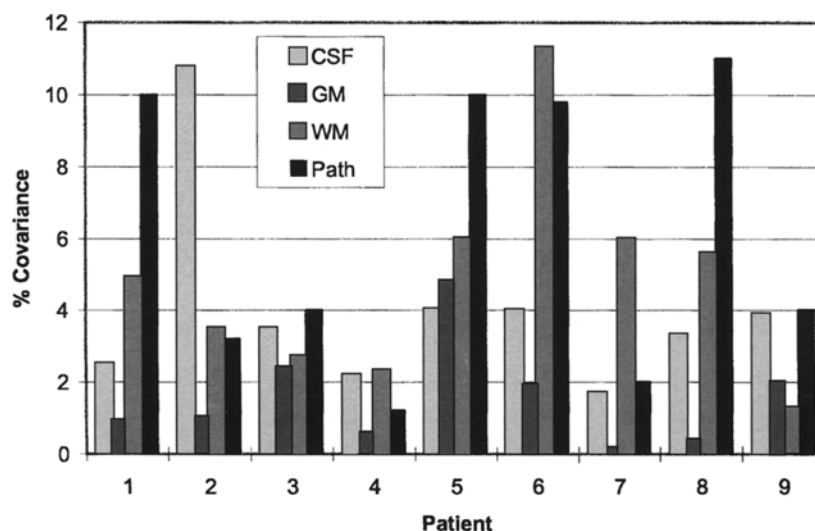
Fig 6. Coefficient of variation of the MS patients scanned twice in a span of 1 week are plotted for GM, WM, CSF, and pathology.

Another approach based on spatial fuzzy connectedness of objects was proposed by Udupa et al,[9] to detect and quantify the volume of MS lesions. In this approach, the operator marks each tissue, and once the algorithm extracts the tissues and pathology the operator is asked to accept or reject each focus of pathology in each slice. All of these methods require significant user time.

There have been efforts to develop totally automated image segmentation techniques. Ardekani et al[10] reported a 3.3% average error for automated detection of intradural tissues from extradural tissues. This is not directly comparable to our results because we report misclassification between intradural tissues, but our errors are of similar magnitude to theirs.

An important advantage of an automated process is the lack of operator variability; the interoperator and intraoperator variability of semiautomatic methods introduces measurement uncertainties. As an example, the complexity of MS lesion identification makes it difficult to obtain consistent results with manual or semiautomatic techniques, and therefore serves as a very challenging pathology for evaluating segmentation reproducibility. In fact, change in an experienced operator's use of a semiautomated technique over the course of a multiyear study was noted in one major study,[1] and if another study design had been used, could have affected the outcome.

There are many reports on MRI quantitation of MS lesions in the literature because it plays an important role in assessing MS therapies.[11-15] Di-

rect comparison is complicated in part because of the variability in how results of a technique is reported. Because the lesion load in MS patients can vary by orders of magnitude, reporting variability as absolute volumes or percent change is probably not useful. It appears that the COV is becoming an accepted measure, because it tends to correct for total volume. Intraobserver variability when using manual methods is now reported in the range of less than 6% for measuring T2 lesion volume (using the same MRI images) and as high as 20% for interobserver variability.[16] Obviously, for a totally automated technique like the one described, intraobserver and interobserver variability is 0.

MS lesions are notoriously difficult to reproducibly segment because they often have ill-defined borders. Because they are multiple and often small, the borders contribute to a significant percentage of total lesion volume. It is not surprising, therefore, that there would be significant variation when humans attempt to separate a lesion from normal tissue and that relative accuracy improves as total lesion volume increases.[17] This might help to explain why intraobserver variability is greater than interobserver variability.[18] We believe that our MS model represents one of the most challenging measures of error that one might expect to see in clinical practice.

Finally, for real-world problems, we must also recognize that variability is introduced by changes in patient position. We have attempted to create a real-world scenario, as we allowed a full week to

pass between scanning, and only routine physical placement methods were used to position patients, as opposed to special scanning-based techniques which have been described.[19] Although some reports have used these special methods, many have reported reproducibility based on a single data set that produces an unrealistically low estimate of variation. We have chosen to use routine clinical positioning techniques because our aim was to develop a technique for mainstream application, not for research protocols. This resulted in a significant percentage change in CSF volume for one patient, resulting from one additional slice through the high cervical region, but this should not represent a problem for routine clinical application.

Several criticisms can be directed at this study. Only one data set was manually segmented to demonstrate accuracy of classification and segmentation of normal structures. There are two reasons for this. First, as noted earlier, the manual technique is subject to great variability, and is a poor gold standard. Second, manual methods are based on visual observation, and this observation of accuracy occurs with every data set the algorithm is applied to. Using this visual test, the algorithm generally performs well. On occasion, portions of facial musculature will remain after morphologic filtering (not observed in the study cases), which is both easily identified, and easily corrected by manual intervention after the algorithm has completed.

The second criticism is that the synthetic lesions in this one data set may not precisely emulate multiple sclerosis (or any other type of pathological) lesions. This is a valid concern, but we would note that in MS, the best gold standard (histopathologic examination of tissues) does not correlate with T2 abnormality either—many lesions are present pathologically in the "normal appearing white matter."[20] Therefore, the best gold standard we could imagine was to create image patterns that visually and numerically simulated real lesions. Knowing what the "normal" voxel was before, and what the voxel value became, we had a reasonable gold standard for determining presence or absence of abnormality.

There is clearly room for improvement in the method. The most obvious improvement would be to implement a final quality control (QC) check by the radiologist to make sure that obvious failures (eg, including extradural tissues) are corrected.

Although not seen in this study, examples where the morphological filtering did not exclude extradural tissues have occurred.

Requiring four image types reduces the usefulness of this algorithm. We are pursuing the possibility of using only T1 and FLAIR images. This would reduce the imaging time required, and also reduce registration demands.

We also believe that because k-NN techniques may model partial volume effects well, its addition could improve reproducibility. Therefore, we are also attempting to further refine the algorithm by applying a neural network based approach to the output of the current classification scheme.

CONCLUSION

In conclusion, we have described and validated a method for fully automated classification and segmentation of head MRIs. Its accuracy has been documented by using manually segmented images with synthetically created lesions of known volume and location. Reproducibility has been shown by repeated scanning of MS patients in a routine clinical fashion. An automated technique such as this may facilitate use in a clinical setting because most (if not all) of the processing has occurred before interpretation. Little time penalty is incurred by the user, and the benefit is a standardized, reproducible identification and measurement of pathological processes.

ACKNOWLEDGMENT

APPENDIX A. THE SEGMENTATION ALGORITHM

1. Initialization—get the image files (sequences to be segmented). These will be referred to as T1, SD, T2, and FLAIR.
2. Spatial registration—geometrically transform the T1, SD, and T2 to match the FLAIR. This corrects for any patient motion between sequences.
3. Intensity Normalization—normalize voxel intensity based on fat.
   (A) Find subcutaneous (scalp) fat. This is done by finding a large group of peripherally located voxels which are at least 75% of the maximum intensity on the T1 weighted images.

(B) For all 4 pulse sequences do{
    1. For each slice of sequence do{
        a. Compute mean intensity of fat for each slice.
        b. Multiply each voxel of each slice by (1000/Mean Fat Intensity for this slice). After this scaling, the mean fat intensity of each slice will be 1000. This also corrects for RF heterogeneity in the Z-direction.
        }end for loop
    } end for loop

****At this point, all voxels of all sequences are normalized to fat intensity****

4. Initial classification. For each voxel do {

if (intensity of voxel on each of the for sequences is compatible with WM),
voxel mask[21] = WM;

if (intensity of voxel on each of the for sequences is compatible with GM),
voxel mask[21] = GM;

if (intensity of voxel on each of the for sequences is compatible with CSF),
    voxel mask[21] = CSF;

    else voxel mask[21] = UNKNOWN;

    } end for loop

****Now there is a large blob of "brain" but also many voxels outside the brain (eg, facial muscles) which are labeled brain. Furthermore, abnormal tissue, like bone and air, are labeled UNKNOWN***

5. Morphologic Filtering

    (A) Do 3D region growing of all groups of brain voxels

    (B) Select the largest group—this is assumed to include the brain

    (C) Apply 3 rounds of erosion and conditional dilation to delete the 'bridges' of brain-like tissue that connect to extracranial tissues.

****Now we have the one large blob of brain, which is disconnected from the face****

6. Restore CSF voxels that are 3D-connected to brain.

7. Label UNKNOWN voxels that are contained within the brain.

    (A) Compute means and standard deviations for WM, GM, and CSF for each of the 4 pulse sequences.

    (B) For each unknown, classify as WM, GM or CSF if the unknown is within 2 standard deviations of the mean on all 4 pulse sequences.

    (C) For any remaining unknowns,
        1. if ((voxel intensity on FLAIR) < mean WM intensity − 2 * SD)
        voxel mask[I] = CSF
        2. if ((voxel intensity on FLAIR) > mean WM intensity + 2 * SD)
        voxel mask[I] = PATHOLOGY
        3. Else assign as WM.

    (D) For each PATHOLOGY voxel, label as ENHANCING if T1 intensity is more than 2SDs above mean WM intensity on T1, assign as NECROTIC is less than mean − 2 * SD), else assign as ISOINTENSE PATHOLOGY.

8. Store the brain mask (segmentation labels).

REFERENCES

1. Paty DW, Li DKB: Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. II. MRI analysis of a multicenter, randomized, double-blind, placebo-controlled trial. Neurology 43:662-667, 1993

2. Simmons A, Arridge AR, Barker GJ, et al: Improvements to the quality of MRI cluster analysis. Magn Reson Imaging 12:1191-1204, 1994

3. Ostuni JL, Levin RL, Frank JA, et al: Correspondence of closest gradient voxels—a robust registration algorithm. J Magn Reson Imaging 7:410-415, 1997

4. Gerig G, Kubler O, Kikinis R, et al: Nonlinear anisotropic filtering of MRI data. IEEE Trans Med Imag 11:221-232, 1992

5. Vannier MS, Speidel CM, Rickman DL: Magnetic resonance imaging multispectral tissue classification. News Physiol Sci 3:148-154, 1988

6. Jackson EF, Narayana PA, Wolinsky JS, et al: Accuracy and reproducibility in volumetric analysis of multiple sclerosis lesions. J Comput Assist Tomogr 17:202-205, 1993

7. Clarke LP, Velthuizen RP, Phuphanich S, et al: MRI: Stability of three supervised segmentation techniques. Magn Reson Imaging 11:95-106, 1993

8. Velthuizen RP, Clarke LP, Phuphanich S, et al: Unsuper-vised measurement of brain tumor volume on MR images. J Magn Reson Imaging 5:594-605, 1995

9. Udupa JK, Wei L, Samarasekera S, et al: Detection and quantification of MS lesions using fuzzy topological principles. Proc SPIE 1996 2710:81-91, 1996

10. Ardekani BA, Braun M, Kanno I, et al: Automatic detection of intradural spaces in MR images. J Comput Assist Tomogr 18:963-969, 1994

11. Paty DW, Li DKB, Oger JF, et al: Magnetic resonance imaging in the evaluation of clinical trials in multiple sclerosis. Ann Neurol 36:S95-S96, 1994 (suppl)

12. Miller DH, Barkhof F, Berry I, et al: Magnetic resonance imaging in monitoring the treatment of multiple sclerosis: concerted action guidelines. J Neurol Neurosurg Psychiatry 54:683-688, 1991

13. Miller DH, Albert PS, Barkhof F, et al: Guidelines for the use of magnetic resonance techniques in monitoring the treatment of multiple sclerosis. US National MS Society Task Force. Ann Neurol 39:6-16, 1996

14. McDonald WI: NMR in diagnosis, monitoring treatment and epidemiology of multiple sclerosis. Acta Neurol Scand 161:52-53, 1995 (suppl)

15. McFarland HF, Frank JA, Albert PS, et al: Using gadolinium-enhanced magnetic resonance imaging lesions to monitor disease activity in multiple sclerosis. Ann Neurol 32:758-766, 1992

16. Paty DW: Magnetic resonance in multiple sclerosis. Curr Opin Neurol 6:202-208, 1993

17. Mitchell JR, Karlik SJ, Lee DH, et al: Computer-assisted identification and quantification of multiple sclerosis lesions in MR imaging volumes in the brain. J Magn Reson Imaging 4:197-208, 1994

18. Narayana PA, Jackson EF, Wolinsky JS: Inter- and intraoperator variability in the quantitative volumetric measure-ments of multiple sclerosis lesions. J Magn Reson Imaging 3:111-124, 1993 (suppl)

19. Galagher HL, MacManus DG, Webb SL, et al: A reproducible repositioning method for serial magnetic reso-nance imaging studies of the brain in treatment trials for multiple sclerosis. J Magn Reson Imaging 7:439-441, 1997

20. Grossman RI: Magnetization transfer in multiple sclero-sis. Ann Neurol 36:S97-S99, 1994

21. Barnes D, Munro PMG, Youl BD, et al: The longstanding MS lesion. A quantitative MRI and electron microscopic study. Brain 114:1271-1280, 1991