Validation Tools for Image Segmentation

Dirk Padfield and James Ross

GE Global Research, Niskayuna, NY, 12309, USA

ABSTRACT

A large variety of image analysis tasks require the segmentation of various regions in an image. For example, segmentation is required to generate accurate models of brain pathology that are important components of modern diagnosis and therapy. While the manual delineation of such structures gives accurate information, the automatic segmentation of regions such as the brain and tumors from such images greatly enhances the speed and repeatability of quantifying such structures. The ubiquitous need for such algorithms has lead to a wide range of image segmentation algorithms with various assumptions, parameters, and robustness. The evaluation of such algorithms is an important step in determining their effectiveness. Therefore, rather than developing new segmentation algorithms, we here describe validation methods for segmentation algorithms. Using similarity metrics comparing the automatic to manual segmentations, we demonstrate methods for optimizing the parameter settings for individual cases and across a collection of datasets using the Design of Experiment framework. We then employ statistical analysis methods to compare the effectiveness of various algorithms. We investigate several region-growing algorithms from the Insight Toolkit and compare their accuracy to that of a separate statistical segmentation algorithm. The segmentation algorithms are used with their optimized parameters to automatically segment the brain and tumor regions in MRI images of 10 patients. The validation tools indicate that none of the ITK algorithms studied are able to outperform with statistical significance the statistical segmentation algorithm although they perform reasonably well considering their simplicity.

Keywords: Validation, Segmentation, Statistical Methods, MRI

1. INTRODUCTION

Segmentation is a critical step for numerous image analysis tasks, and a large number of segmentation algorithms have been developed over the years. The software implementation of these algorithms often entails exposure of parameters to the user so that algorithm behavior can be customized for specific applications. However, in practice it can be difficult to determine optimal parameter settings for a given task; furthermore, it is not always clear which algorithm should be preferred. The purpose of our work is to develop tools for optimizing segmentation parameter settings for a single case and across a collection of cases, and for using statistical analysis to compare the effectiveness of various segmentation algorithms. To demonstrate these tools, we evaluate several region-growing segmentation algorithms of the Insight Toolkit,[1] and compare their effectiveness with a statistical segmentation algorithm developed by the Brigham and Women's Surgical Planning Lab (SPL).[2] The validation methodology of this paper will hopefully serve as a guide for researchers to effectively apply segmentation algorithms for their specific image analysis tasks. In Section 2 we present the methods of our study, in Section 3 we present our results, and we draw our conclusions in Section 4.

2. METHODS

In this section, we first describe the data of the experiments. We then outline the various ITK region-growing algorithm under investigation as well as the placement of the seeds and the post-processing steps. This is followed by a description of the parameter optimization stage including metrics and the use of the Design of Experiments framework for parameter refinement. The section is concluded with a discussion on statistical analysis of the algorithm performance.

Further author information: (Send correspondence to Dirk Padfield.)
Dirk Padfield: E-mail: padfield@research.ge.com

2.1 Data

For our study, we used ten MR images taken from the Brigham and Women's Tumor Database. The database itself consists of MR images of twenty anonymous brain tumor patients, as well as segmentations of the brain and tumor from these scans (manual segmentations obtained by neurosurgeons and automated segmentations obtained by the SPL method). The SPL group trained their algorithm on the first ten cases and used all twenty cases for validation. In each case, one of the following tumor types is present: meningioma, low grade glioma, or astrocytoma. Table 1 presents the tumor types and their anatomical locations for the ten cases investigated in this validation study.

According to,[2] the patients' heads were imaged in the sagittal plain with a 1.5T MRI system (Signa, GE Medical Systems, Milwaukee, WI), with a post contrast 3D sagittal spoiled gradient recalled (SPGR) acquisition with contiguous slices (flip angle, 45 degrees; repetition time (TR), 35 ms; echo time (TE), 7 ms; field of view, 240 mm; slice-thickness, 1.5mm; slice gap, 0.0mm; 256x256x124 matrix). The data is presented as 256x256 16 bit data, with the pixel size being 0.9375 x 0.9375 mm.

The ground truth used in this study was the same as the ground truth described in.[2] Their ground truth segmentations were based on manual segmentations using interactive computer segmentation tools. To minimize the influence of inter-observer variability and human error, the ground truth was based on the segmentations of four independent human observers. A single 2D slice was randomly selected from the subset of the MRI volume that showed the tumor. On this slice the human observers manually outlined the brain and tumor. The ground truth segmentation of brain and tumor in each patient dataset was defined as the set of pixels where at least three out of four observers agreed on the identification. All other pixels were labeled as background.

2.2 Algorithms

We evaluated three region-growing segmentation algorithms from the Insight Toolkit: Threshold Connected (TC), Neighborhood Connected (NC), and Confidence Connected (CC). Each of these algorithms proceeds by considering the neighbors of foreground pixels, each time using an intensity interval and spatial continuity to determine how a pixel should be labeled. We compared the accuracy of these algorithms with the one developed by the SPL,[2–4] which is a semi-automated process that relies on explicit anatomical information derived from a digital atlas. The technique requires a user to select three to four example points for each tissue class, and the program then uses these points to calculate statistical models of the tissue distributions. The anatomical atlas is then used to guide hierarchical statistical classification of the various tissue types.

The TC segmentation algorithm uses an intensity interval and spatial continuity about a seed point in order to define a structure. It begins by considering the intensity value of a user-specified seed point. If the intensity of that point is within a predefined intensity interval, the point is labeled foreground. The algorithm proceeds by considering the neighbors of foreground pixels, each time using the interval to determine how a pixel should be labeled. This process continues until no more foreground neighbors are found.

Table 1. Tumor types and locations for the cases considered in this study.

Case	Tumor	Location
1	Meningioma	Left Frontal
2	Meningioma	Left Parasellar
3	Meningioma	Right Parietal
4	Low Grade Glioma	Left Frontal
5	Astrocytoma	Right Frontal
6	Low Grade Glioma	Right Frontal
7	Astrocytoma	Right Frontal
8	Astrocytoma	Left Temporal
9	Astrocytoma	Left Frontotemporal
10	Low Grade Glioma	Left Temporal

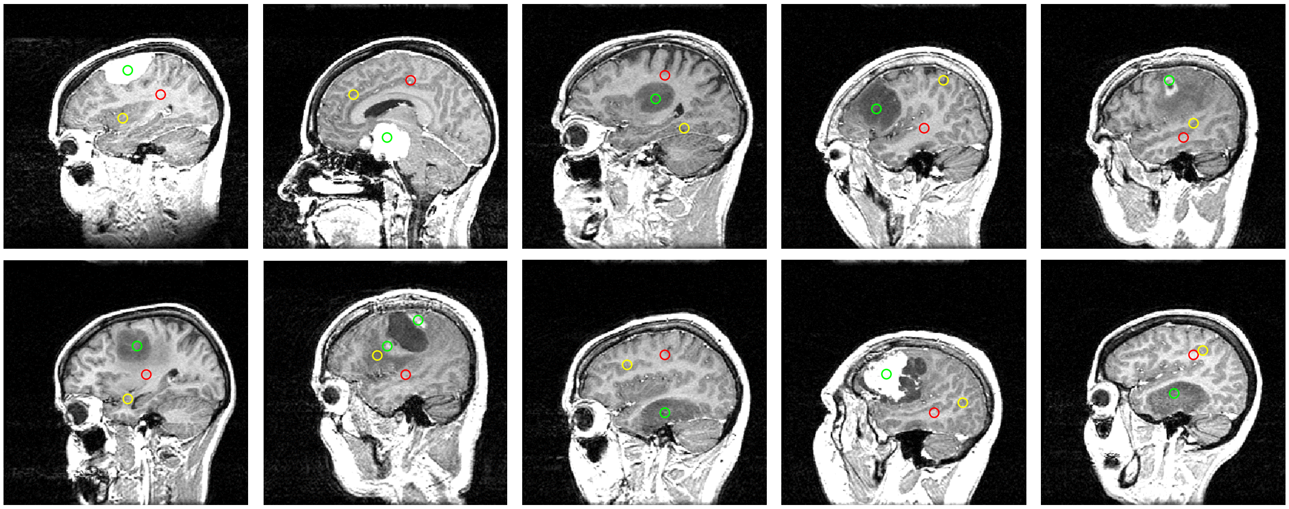

Figure 1. Brain and tumor seed locations for the ten cases in our study. Green indicates the tumor seed, yellow indicates the gray matter seed, and red indicates the white matter seed.

The NC segmentation algorithm is similar to the TC algorithm, but a given pixel is labeled foreground only if all its neighbors also have intensity values that fall within the pre-defined interval; this criterion makes ”leaking” less likely. The number of neighbors here depends on a user-defined radius value.

The CC algorithm also begins segmentation by considering a user-specified seed point. However, rather than using a prescribed intensity interval, a confidence interval is constructed from the statistics computed in a neighborhood about the seed point. The algorithm computes the mean and standard deviation of the pixel values currently labeled as foreground and uses a term to multiply the standard deviation in order to define a new interval. All neighboring pixels within this interval are labeled foreground. When no more neighbors are found within this interval, the mean and standard deviation are updated, and the process is repeated for a given number of iterations, which is another parameter.

2.3 Seeds

The ITK region-growing algorithms discussed here require seed points to be specified. For our validation experiments, we set two seed points in the brain region: one in the white matter and one in the gray matter. For both the TC and NC algorithms, we took the union of the segmentations corresponding to these seeds to yield the whole brain segmentation. For the CC algorithm, we generated segmentations for each seed as well as a segmentation using the statistics of both seeds. We then took the union of these segmentations. For the tumor region, we only set one seed point (case seven is an exception, since two tumors are present). Figure 1 displays the seed point locations on each of the case slices.

2.4 Post Processing

The binary images resulting from the region-growing segmentations are inherently noisy because of the noise of the original images, and excessive pre-smoothing is undesirable. Therefore, we carried out a post-processing step to fill in the holes of the segmented object. In this process, we took care not to fill the tumor cavity since tumors are generally within the brain region. To remove the noise while retaining larger cavities, we performed morphological closing on the images. We empirically chose the radius of the closing operation to be four pixels for both the CT and the CC segmentations. However, since the NC algorithm forces all pixels within a neighborhood to be within the threshold values in order for a pixel to be considered foreground, the resulting images have unnatural-looking square blocks with sides equal to twice the radius (which was set at two pixels). Because this also occurs on the edges of the object, the resulting object is under-segmented by an amount equal to the radius. Therefore, in this case, the closing operation was composed of dilation by four pixels and erosion by two pixels (instead of four), which filled in many of these blocks and dilated the edge to its correct location. Figure 2 shows a pre- and post-processed Neighborhood Connected segmentation.

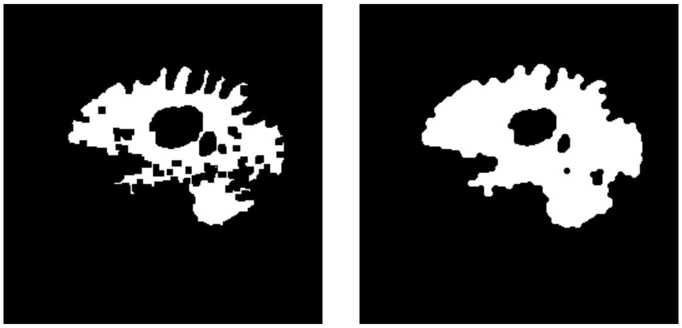

Figure 2. Neighborhood Connected result showing the raw result on the left and the output of post-processing on the right.

2.5 Parameter Optimization

Each of the ITK algorithms investigated in this validation study has several adjustable parameters whose values impact the accuracy of the segmentation. For the TC algorithm, we varied both the upper and lower threshold values. The same was done for the NC algorithm, and the neighborhood radius was kept constant at two. For the CC algorithm, we varied both the multiplier and number of iterations.The goal of this study was to find the parameter settings for each algorithm that maximized the value of the performance metric.

To compare the effectiveness of the algorithms, it is necessary to find the parameter settings for each algorithm that maximizes the value of a similarity metric. Here we use a similarity index, S, between two sets A and B

$$S = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

where $||$ represents the size of the set, and $\cap$ represents the intersection of the two sets. The sets A and B in this case are the sets of segmented pixels extracted by the segmentation algorithms and ground truth, respectively. This index takes into account the size of the overlap, but it does not depend on image size; it ranges from 0 (no overlap) to 1 (perfect alignment).[5]

A recursive Design of Experiment (DOE) approach is used to search for the optimal parameter settings for a given algorithm. First, we perform a sparse sampling of the parameter space to generate an optimization surface based on the accuracy measured from Equation 1. In successive iterations, a more focused search is carried out on the range of values of the optimization surface for which the accuracy is highest. For example, in the case of segmenting the brain using the Connected Threshold algorithm, the maximum brain intensity is around 300, so the settings for the lower and upper thresholds were first set as: Lower Threshold $LT = 0 : 25 : 300$, Upper Threshold $UT = 5 : 25 : 305$. The notation here means that, for example, the lower threshold was set between 0 and 300 with an increment of 25. Using this range, it was found that the optimal LT was between 50 and 75, and the optimal UT was between 105 and 130. Therefore, the range was refined to: $LT = 40 : 5 : 85$, $UT = 95 : 5 : 140$. With this new range, the optimal LT was found to be 65, and the optimal UT to be 120. Thus, the range was refined to: $LT = 50 : 1 : 80$, $UT = 105 : 1 : 135$. This search refinement continues until a negligible increase in accuracy is realized. An example optimization surface is given in Figure 3, which demonstrates the effect of the parameter values on segmentation accuracy for a particular algorithm and case.

The optimal parameter settings for a given algorithm and a given case are found as the settings that maximize the accuracy metric S. Thus, for a single case, the vector of optimal parameter values, $\overline{p}*$, is found as

$$\overline{p}* = \arg\max_{\overline{p}} \frac{2|A_{\overline{p}} \cap B|}{|A_{\overline{p}}| + |B|} \tag{2}$$

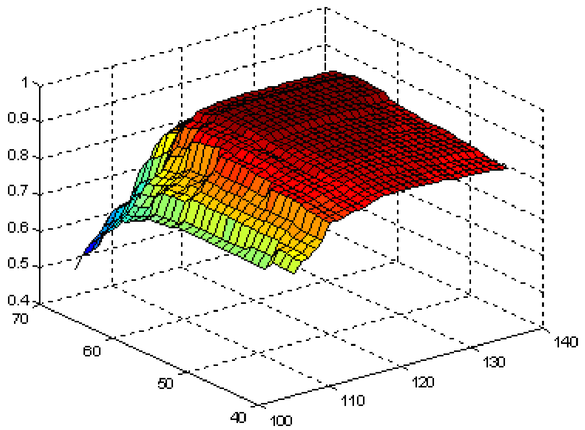

Figure 3. Optimization surface. As the values of the parameters are varied, the segmentation accuracy changes.

where $\overline{p}$ is an $d \times 1$ dimensional vector, with each of the d elements corresponding to a parameter setting under investigation; $A_{\overline{p}}$ is the output segmentation corresponding to the specific vector values of $\overline{p}$; and B represents the ground truth segmentation for that case.

In general, different cases have different optimal parameter settings. To find parameter settings that are optimal across a set of cases, these case accuracy values need to be combined. One method is to average the accuracy values for each parameter setting across cases

$$\overline{p}*_{ave} = \arg\max_{\overline{p}} \frac{1}{n} \sum_{i}^{n} \frac{2|A_{\overline{p},i} \cap B_i|}{|A_{\overline{p},i}| + |B_i|} \tag{3}$$

where n is the number of cases and $\overline{p}*_{ave}$ is the $d \times 1$ vector containing the optimal parameter values taken over all cases. Combining accuracy values in this way results in an equal weighting of each of the case scores. However, if one case has small structures and another has large structures but both have approximately equal accuracy values, they will influence the overall accuracy equally. To overcome this, the parameter settings can be optimized using the following equation:

$$\overline{p}*_{ave} = \arg\max_{\overline{p}} \frac{\sum_{i}^{n} 2|A_{\overline{p},i} \cap B_i|}{\sum_{i}^{n} |A_{\overline{p},i}| + |B_i|} \tag{4}$$

Here, if one case has small structures and another has large structures, the large structure will have a greater weight in the average. Therefore, this is a more accurate equation for finding the optimized parameters.

2.6 Statistical Analysis of Algorithm Performance

In order to predict how accurate the segmentation algorithms will be on novel data, we employed a leave-one-out strategy. The cases were rotated such that a single case was withheld, and the remaining cases were used for training. We found the optimal parameter settings for the average of those cases, and then tested these parameter settings on the case that was originally excluded, providing a measure of the algorithm's effectiveness on new data.

To test whether the ITK segmentation algorithms are as accurate as the SPL algorithm, we used a measurement framework employing aspects of statistical analysis. It can be said with 95% confidence that an algorithm produces more accurate segmentations than another if the average accuracy over the cases is higher than that of the other algorithm and at least 80% of the cases have higher accuracy values than the other algorithm. The results of the comparisons are given in the Results section.

Table 2. Optimal settings for each of the algorithms. Only one set of parameters is needed for the brain segmentations, but three sets are needed for the tumor segmentations, depending on whether the tumors are bright, middle, or dark. For the CT and NC algorithms, the optimal lower threshold (LT) and upper threshold (UT) parameters are given. For the CC algorithm, the optimal multiplier (M) and iterations (I) parameters are given.

Algorithm	Brain	Tumor (bright)	Tumor (middle)	Tumor (dark)
CT	LT = 69, UT = 104	LT = 115, UT = 210	LT = 80, UT = 245	LT = 35, UT = 70
NC	LT = 61, UT = 122	LT = 110, UT = 255	LT = 70, UT = 245	LT = 35, UT = 80
CC	M = 1.9, I = 5	M = 2.7, I = 2	M = 2.1, I = 2	M = 2.2, I = 2

3. RESULTS

3.1 Parameter Settings

The optimal parameter settings for each of the algorithms are given in Table 2. These settings were found using the DOEs with the leave-one-out strategy with Equation 4. One pair of settings was used for the brain images, but three pairs were used for the tumors since their intensities fit into the three categories of bright, middle, and dark. The brain segmentation results given these settings are shown in Figure 4, and the tumor segmentation results are shown in Figure 5.

3.2 Accuracy Results

The brain accuracy results are summarized on the left in Table 3. Only the Neighborhood Connected algorithm has accuracy greater than the SPL segmentations, and none of the algorithms meet the requirement that 80% of the cases must have higher accuracy. The tumor accuracy results are summarized on the right in Table 3. None of the algorithms have greater accuracy than the SPL segmentations, and none of the algorithms meet the requirement that 80% of the cases must have higher accuracy. The Neighborhood Connected algorithm has the lowest accuracy values among the tumor images.In fact, using the statistical test of Section 2.6 shows that the only segmentations that can be said to be better are the SPL over the NC for the tumor images. Of the ITK algorithms, the Confidence Connected algorithm performs best for combined brain and tumor segmentations. This is intuitive since this algorithm adapts its statistics based on the included pixels as it grows, whereas the CT and NC algorithms depend on absolute upper and lower threshold values. Although the ITK algorithms are relatively simplistic, the accuracy results are quite close to those of the SPL method.

Table 3. Summary of brain and tumor ITK segmentation results as compared to the SPL segmentation results. CT is the Connected Threshold, NC is the Neighborhood Connected, and CC is the Confidence Connected algorithm.

Case	Brain Segmentation Results				Tumor Segmentation Results			
	CT	NC	CC	SPL	CT	NC	CC	SPL
1	0.85	0.97	0.96	0.95	0.94	0.97	0.97	0.98
2	0.96	0.96	0.96	0.92	0.94	0.91	0.86	0.91
3	0.94	0.93	0.93	0.93	0.97	0.95	0.97	0.96
4	0.91	0.91	0.85	0.93	0.87	0.58	0.88	0.91
5	0.87	0.85	0.84	0.88	0.68	0.71	0.70	0.85
6	0.79	0.95	0.87	0.94	0.96	0.94	0.87	0.98
7	0.86	0.86	0.85	0.82	0.57	0.69	0.93	0.84
8	0.95	0.96	0.96	0.97	0.98	0.96	0.96	0.88
9	0.95	0.96	0.94	0.97	0.93	0.94	0.94	0.97
10	0.91	0.95	0.95	0.89	0.92	0.94	0.94	0.96
Mean	0.90	0.93	0.91	0.92	0.87	0.86	0.90	0.93
Std.	0.06	0.04	0.05	0.05	0.14	0.14	0.08	0.05
# > SPL	4	5	5	N/A	3	1	3	N/A

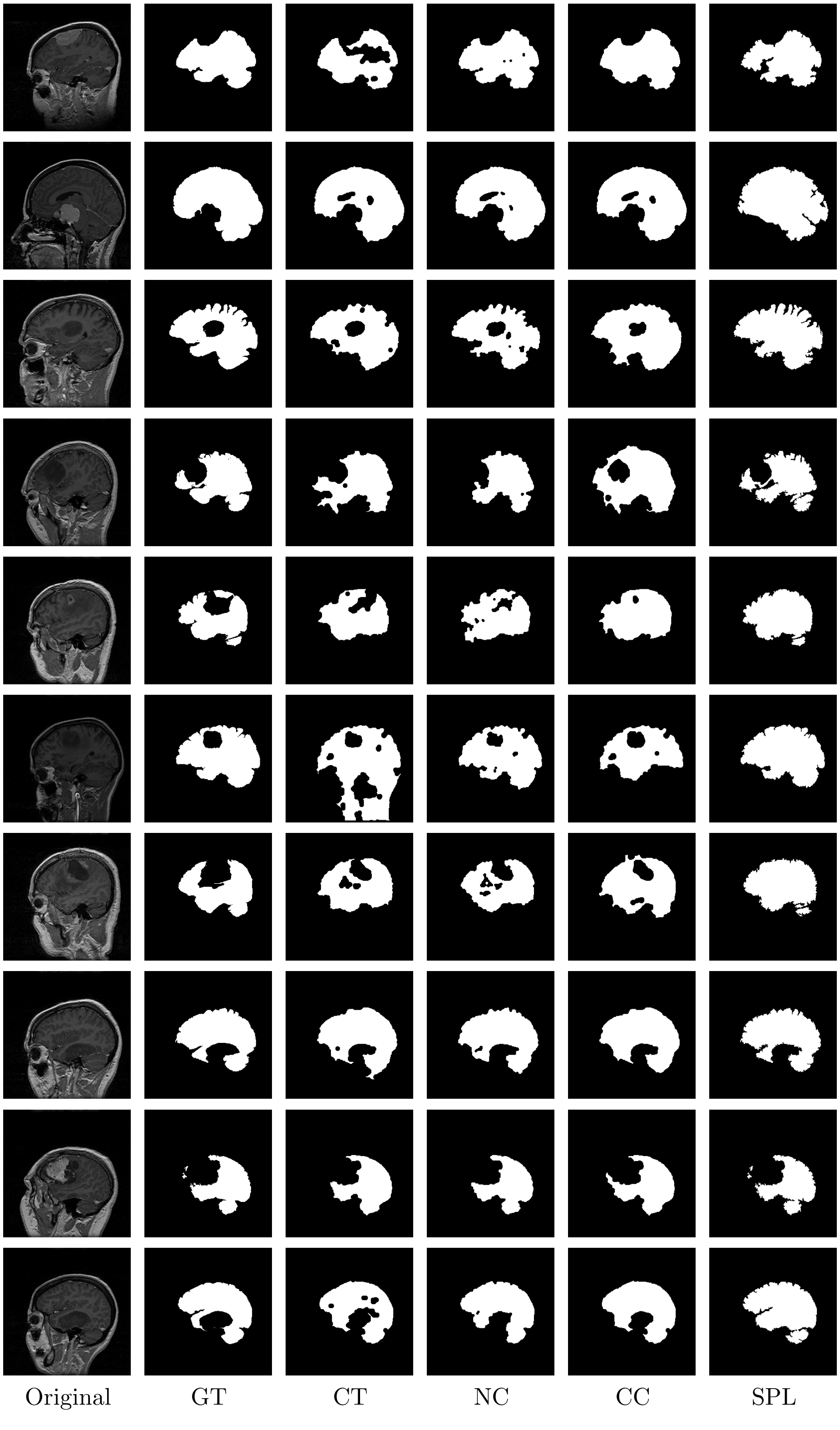

Figure 4. Comparison of brain segmentation results for the ten datasets. The first column shows the original images, and the next five show the segmentations for the Ground Truth, Connected Threshold, Neighborhood Connected, Confidence Connected, and SPL algorithms, respectively. The ten rows correspond to the ten cases.

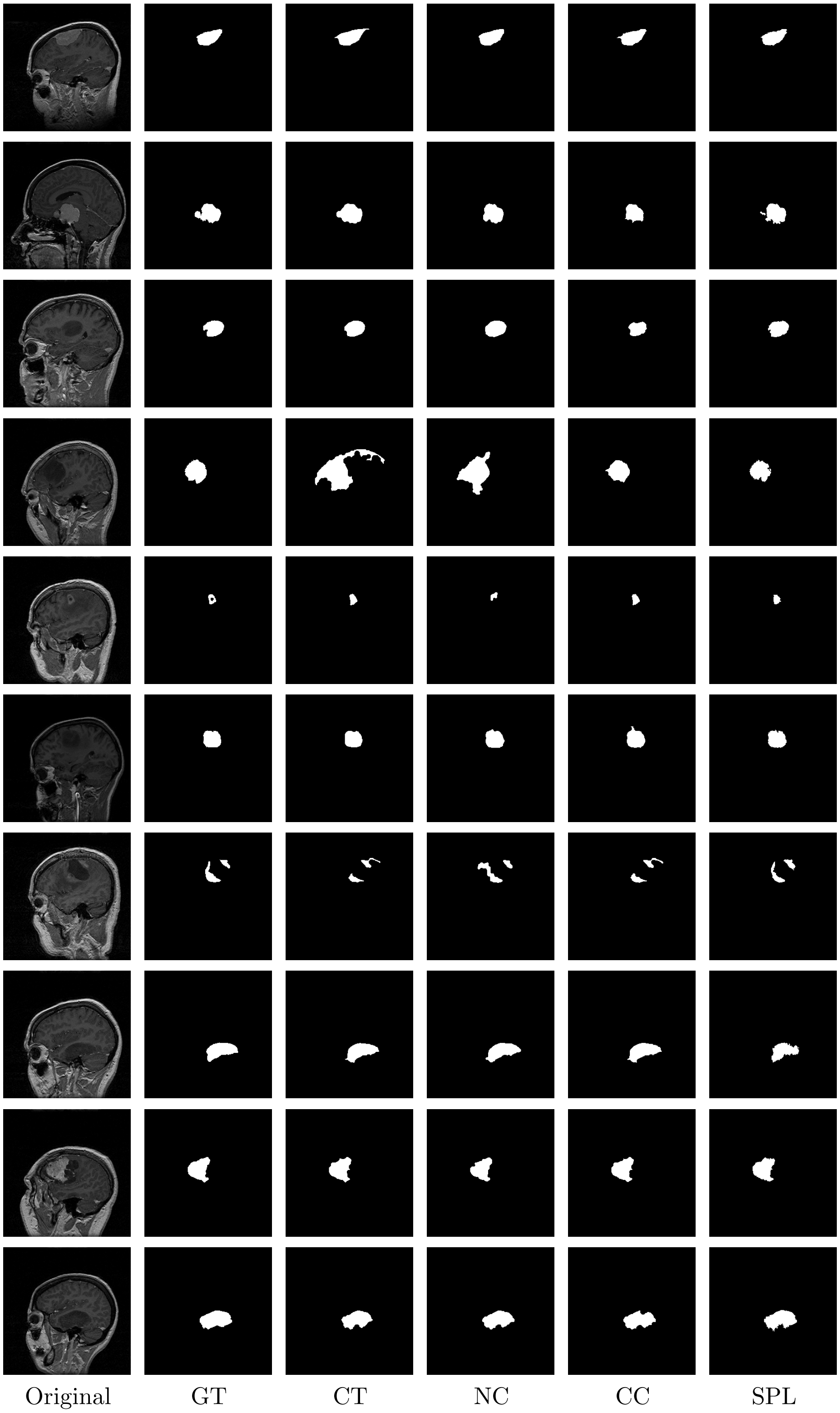

Figure 5. Comparison of tumor segmentation results for the ten datasets. The first column shows the original images, and the next five show the segmentations for the Ground Truth, Connected Threshold, Neighborhood Connected, Confidence Connected, and SPL algorithms, respectively. The ten rows correspond to the ten cases.

4. CONCLUSIONS

We described methods for determining the optimal parameter settings for segmentation algorithms based on measurement accuracy values calculated through comparison with manual segmentations. These optimal parameters are found using a recursive DOE framework and were used to segment both brain and tumor structures. To compare the results of these algorithms with another algorithm, statistical analysis methods were used. The results demonstrated that, although the segmentation accuracy values of the ITK segmentation algorithms were high, they were not statistically more accurate than the SPL segmentations overall. In addition, the results demonstrated that, of the ITK algorithms, the Confidence Connected algorithm achieved the highest accuracy overall for brain and tumor segmentations.

5. ACKNOWLEDGMENTS

This work is part of the Visible Humans Project - Image Processing Tools, funded through the National Library of Medicine, National Institutes of Health, Grant N01-LM-9-3531.

REFERENCES

[1] Ibanez, L., Schroeder, W., Ng, L., and Cates, J., *The ITK Software Guide*. Kitware, Inc. ISBN 1-930934-15-7, http://www.itk.org/ItkSoftwareGuide.pdf, second ed. (2005).

[2] Kaus, M., Warfield, S., Nabavi, A., Black, P., Jolesz, F., and R., K., "Automated segmentation of MRI of brain tumors," *Radiology* **218**(2), 586–591 (2001).

[3] Warfield, S., Kaus, M., Jolesz, F., and Kikinis, R., "Adaptive template moderated spatially varying statistical classification.," in [*MICCAI*], 231–238 (1998).

[4] Warfield, S. K., Kaus, M., Jolesz, F. A., and Kikinis, R., "Adaptive, template moderated, spatially varying statistical classification," *Medical Image Analysis* , 43–55 (2000).

[5] Pathak, S., "Validation study: Segmentation of gray and white matter from T1-weighted MRI images and MS lesion segmentation by multichannel tissue classification," *Radiology* (2002).