Medical Physics
The International Journal of Medical Physics Research and Practice

Ensemble segmentation for GBM brain tumors on MR images using confidence-based averaging

Jing Huo, Kazunori Okada, Eva M. van Rikxoort, Hyun J. Kim, Jeffry R. Alger, Whitney B. Pope, Jonathan G. Goldin, and Matthew S. Brown

Citation: Medical Physics **40**, 093502 (2013); doi: 10.1118/1.4817475
View online: http://dx.doi.org/10.1118/1.4817475
View Table of Contents: http://scitation.aip.org/content/aapm/journal/medphys/40/9?ver=pdfcov
Published by the American Association of Physicists in Medicine

Ensemble segmentation for GBM brain tumors on MR images using confidence-based averaging

Jing Huo[a)]
TeraRecon Inc., 4000 East 3rd Avenue, Suite 200, Foster City, California 94404

Kazunori Okada
Department of Computer Science, San Francisco State University, Thornton Hall 911, 1600 Holloway Avenue, San Francisco, California 94132-4163

Eva M. van Rikxoort
Department of Radiology of Radboud University Nijmegen Medical Centre, Radboud University Nijmegen, 6525 GA Nijmegen, The Netherlands

Hyun J. Kim, Jeffry R. Alger, Whitney B. Pope, Jonathan G. Goldin, and Matthew S. Brown
Department of Radiological Sciences, University of California in Los Angeles, 924 Westwood Boulevard, Suite 650, Los Angeles, California 90024

(Received 17 February 2013; revised 11 May 2013; accepted for publication 3 July 2013; published 20 August 2013)

Purpose: Ensemble segmentation methods combine the segmentation results of individual methods into a final one, with the goal of achieving greater robustness and accuracy. The goal of this study was to develop an ensemble segmentation framework for glioblastoma multiforme tumors on single-channel T1w postcontrast magnetic resonance images.
Methods: Three base methods were evaluated in the framework: fuzzy connectedness, GrowCut, and voxel classification using support vector machine. A confidence map averaging (CMA) method was used as the ensemble rule.
Results: The performance is evaluated on a comprehensive dataset of 46 cases including different tumor appearances. The accuracy of the segmentation result was evaluated using the F_1-measure between the semiautomated segmentation result and the ground truth.
Conclusions: The results showed that the CMA ensemble result statistically approximates the best segmentation result of all the base methods for each case. [http://dx.doi.org/10.1118/1.4817475]

Key words: GBM, ensemble

1. INTRODUCTION

Glioblastoma multiforme (GBM), a World Health Organization (WHO) grade IV astrocytoma, is the most common human brain tumor comprising about 12%–15% of all primary central nervous system (CNS) tumors and accounting for about 50%–60% of all astrocytomas.[1] Survival for patients with glioblastoma, although individually variable, averages 14 months after diagnosis.[2] Clinical trials are investigating effective treatments for GBM brain tumors, and imaging is playing an important role. Contrast-enhanced tumor size change on serial imaging studies is used as a surrogate endpoint using 1D and 2D diameters. Computer-aided volumetric methods are also under investigation, which can be more effective than diameters when the tumor contains a nonenhanced core or has an irregular shape. In clinical studies, manual contouring has been used to segment tumors on MR images. For example, in a recent clinical study of correlating methylated-DNA-protein-cysteine methyltransferase (MGMT) promoter methylation and imaging features of GBM tumors, Drabycz *et al.*[3] used manual contouring for GBM brain tumor segmentation. An accurate and robust automated segmentation system would facilitate quantitative analysis in clinical studies. In Fig. 1, we show a 2D slice of a T1 weighted postcontrast magnetic resonance (MR) image presenting an enhancing GBM brain tumor with the outline of the active tumor region.

Automatic GBM brain tumor segmentation is a challenging task, since brain tumors are heterogenous, and highly variable in size, location, shape, and appearance. They also often deform adjacent structures in the brain. Some artifacts of MR imaging also increase the difficulty of tumor segmentation. Imperfection of the RF pulses and the location of RF coils may introduce nonuniformity in MR images. This study focuses on recurrent GBM brain tumors that develop after surgery, many of which contain a cavity, and the enhancing portions can vary in shape, for example, ring-shape, blob-like shape, or multiple components attached to the cavity or dispersed into the brain tissue (see Fig. 2). Furthermore, when patients are scanned at multiple centers, with different scanners and contrast agent injection protocols, the image intensity contrast can vary greatly. These factors make GBM brain tumor segmentation a very challenging problem in a clinical setting, and there is a lack of studies evaluating GBM brain tumor segmentation methods in a large clinical dataset.

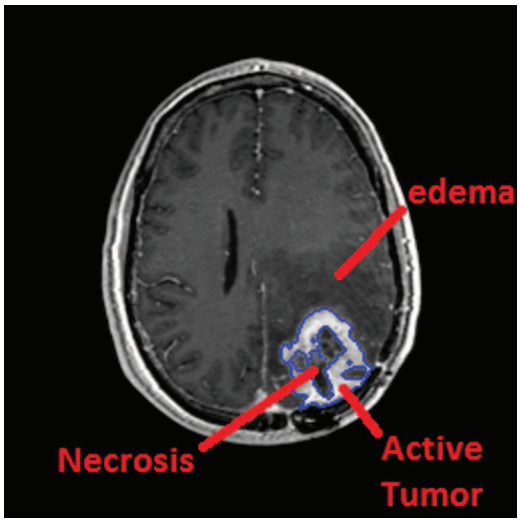

FIG. 1. Example of a GBM brain tumor on a T1w postcontrast MR image slice with the active part contoured.

Computer-based brain tumor segmentation has remained largely experimental work. Many efforts have exploited MRI's multidimensional data capability through multispectral analysis.[4–9] There are generally several categories of techniques: knowledge-based, clustering, voxel-based classification, level set method, and graph-based techniques.

Knowledge-based segmentation systems typically use a brain atlas to provide prior information. Fletcher-Heath *et al.*[10] applied a knowledge-based system to segment nonenhancing tumors. Prastawa *et al.*[11] applied outlier detection to find abnormal regions, applied k-means clustering ($k = 2$) to separate tumor and edema, and then a region competition method using level-sets to add a smoothness constraint. In the study, they used T1-weighted precontrast and T2-weighted images, without contrast injection. However, in clinical trials, tumor definition is based on T1-weighted postcontrast images. They reported that the intrareader variability could be as low as 59.4%.

Among the clustering techniques, fuzzy clustering methods are the approach most widely employed across all tumor types. Fuzzy C-means (FCM) clustering is used frequently, since it does not require training data. Phillip *et al.*[4] was the first to apply FCM clustering to GBM brain tumor segmentation, and correlated the segmentation with tumor histology. The limitation of the study is that it did not include a quantitative validation of the method. Beevi and Sathik[12] applied an efficient denoising algorithm before FCM and incorporated spatial probability to deal with the sensitivity to noise. The limitation of the study is that the method was validated on one clinical brain MR scan with unknown tumor type. Khotanlou *et al.*[13] performed symmetry analysis and fuzzy clustering to initialize the segmentation, and combined deformable model and spatial relations to refine it. It was not clear whether the method was evaluated on GBM tumors, and it would be interesting to evaluate the method on images from GBM clinical trials. Aside from FCM, Ahmed and Mohamad[14] performed k-means clustering combined with the anisotropic diffusion denoising and evaluated on one MR scan. Liu *et al.*[9] developed a semiautomated system using the fuzzy-connectedness method and evaluated the overall volume accuracies for 20 patients. The method requires additional steps to remove attached brain structures. Clark *et al.*[5] use a knowledge-based system including five stages, using T1-weighted, T2-weighted, and PD-weighted image intensity. In each stage, various heuristic parameters are applied. The performance is reported as a correspondence ratio that ranges from 0.43 to 0.85 in 16 scans from seven patients. They used 17 slices from three patients to set up the heuristic parameters. It is not clear how difficult and practical it would be to set universal parameter values in the setting of a large clinical trial, considering the variability of GBM tumors.

Voxel-based supervised classification methods have been investigated by a number of researchers.[15] Vinitski *et al.*[16] developed a system using a k-nearest neighbor classifier (kNN) to segment multiple sclerosis (MS) lesions and brain tumors from a limited number of patients. Validation with more tumor cases is needed to apply the method in clinical trials. Jolesz and co-workers[17] developed an adaptive template-moderated (ATM) classification algorithm (ATS) which incorporated a brain atlas to include spatial anatomical

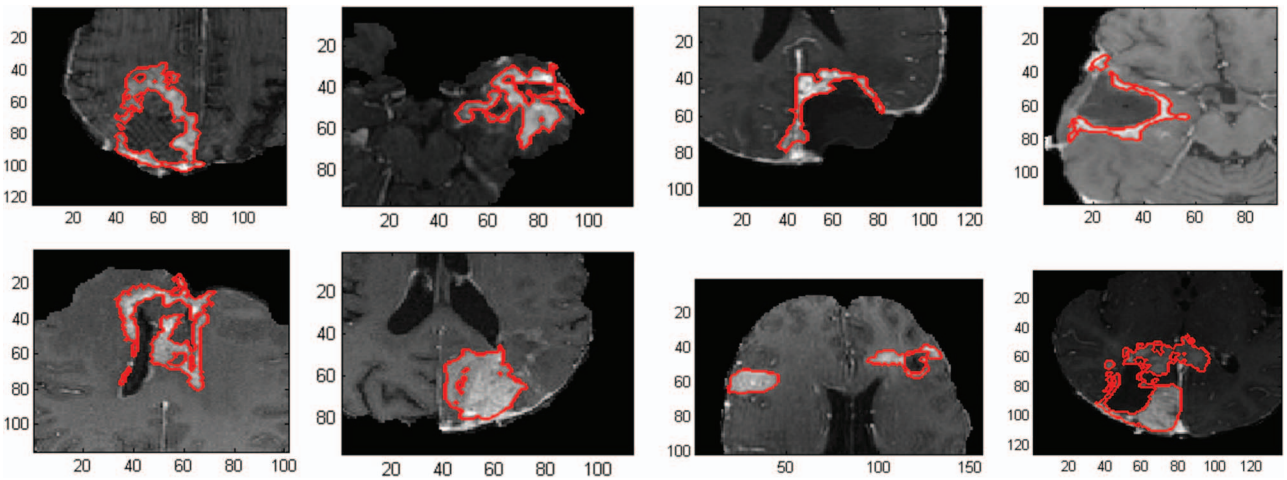

FIG. 2. Examples of recurrent GBM tumors on MR images. The contours show the doctor's reading.

information into the kNN classification system to segment the MR image into five different tissue classes: background, skin, brain, ventricles, and tumor. Kaus *et al.*[18] applied the algorithm to low-grade glioma and meningioma; however, it is not clear how the ATM algorithm will perform for GBM tumors. Prastawa *et al.*[19] applied a system derived by Van Leemput *et al.*[20] to GBM tumor segmentation. The system used the difference between T1w pre and postcontrast images to develop tumor and edema priors, and form a Gaussian mixture model framework solved by expectation-maximization (EM) technique. The performance is reported as an overlap ratio of 0.49–0.92 from five patients. The system was extended for GBM tumor segmentation by adding the tumor and edema classes.[21] One limitation of the study is that they did not provide *a prior* in the model for necrosis, cyst, or cavity. It is common for GBM tumors to have necrosis or a surgical cavity, especially in recurrent GBM. Another limitation of the study is that the simplified geometric model for tumor shape cannot cope with tumors that have complex appearance and poorly defined boundaries. Zhang *et al.*[22] used baseline as training and follow-up as testing images. The method was tested on five scans on one tumor case. The application is limited since the GBM tumors on the baseline images still need to be manually contoured. Schmidt *et al.*[23] developed alignment-based features including a spatial prior, symmetry, intensity, and multiscale texture. The dataset included ten patients with one cavitated tumor from two sites. They reported average overlap of 0.732. However, performance for the active tumor volume is not clear. Lee *et al.*[24] applied discriminative random fields (DRFs) model with a support vector machine (SVM) and reported performance of 0.53–0.89 overlap ratio for 12 scans from seven patients. The weakness of the study is that they used patient-specific training, which means training and testing voxels are from the same patient, and the manual contouring is still needed for each patient. Ayachi and Amor[25] applied a support vector machine (SVM), using nine slices from each tumor as training, and the rest of the slices on the same patient as testing, and report a 0.82 true positive rate for four cases. However, with patient-specific training, manual contouring is still needed. Zhang *et al.*[26] applied a multikernel SVM, and again the limitation of the study is the need for patient-specific training.

Level set and graph-based methods have also been explored for brain tumor segmentation. Ho *et al.*[6] ran a level set algorithm on probabilities derived from a T1w pre and postcontrast difference image. They report an 80%–90% overlap ratio on three tumors of blob-like shape. However, it is not clear how the method performs for irregular tumor shapes. Popuri *et al.*[27] extracted a clustered feature set, integrated them into a level set framework and used a Dirichlet prior to exclude the surrounding tissues. They showed success in differentiating tumor from normal tissue by incorporating shape information; however, it is not clear how it performs for GBM tumors which usually have irregular shapes. Taheri *et al.*[28] used a threshold-based speed function for level-set function evolution. Corso *et al.*[7,29] developed a segmentation by weighted aggregation (SWA) algorithm based on graph shift algorithm for GBM brain tumor segmentation. Dube *et al.*[8] incorporated the texture features into the SWA framework and applied to the GBM brain tumor segmentation on one-channel MRI using T1-weight postcontrast MRI. The study achieved 70% accuracy for the majority of the cases; however, the failure cases will need to be addressed before it is ready for the clinic. Recently, other features other than intensity were studied, including grayscale concurrence matrix (GLCM) features,[30] discrete cosine transform (DCT) features,[31] and the Gabor wavelet filter.[32]

In summary, most of the literature reports the use of multichannel MR to segment GBM tumors, while segmentation on a single-channel MR has only been reported infrequently.[8] Although multichannel MR sequences are useful in differentiating brain tissues and disease, they are usually acquired at low resolutions, with slice gaps, and images from different sequences are often not aligned. Images can be realigned to a reference series but the resliced image series can suffer from lower resolutions along the slice axis as well as slice gaps. It is now possible to perform high resolution 3D imaging using various contrast mechanisms (T1w, T2w, FLAIR) and using identical image parameters for each image set on modern MR scanners. However, even with same-resolution T2w or FLAIR scans, they are not scanned at the same time, and registration is still needed to align them. That might be a source of errors. Due to the time issue, it is not standard care to acquire high-resolution for T2w and FLAIR images in the current clinical practice. Segmentation on a single channel T1 postcontrast isotropic data is potentially important in determining tumor volume for therapeutic response assessment in clinical trials.

Most of the papers reported a small dataset of less than ten cases to evaluate their methods. It is not clear whether the techniques can handle the more difficult and irregular GBM tumors that inevitable arise in larger clinical datasets.

There is also limited investigation of irregular recurrent GBM tumors. The tumor recurrence could happen around the surgical cavity or at a distant site, and show diffuse-pattern with anti-VEGF drugs. These factors increase the difficulty of recurrent GBM tumor segmentation compared to the newly diagnosed GBM tumors.

The contribution of this study is to investigate an ensemble approach to GBM tumor segmentation that combines results from three individual general-purpose segmentation algorithms, aiming to achieve high accuracy in GBM tumor segmentation.

There has been active research on combining multiple segmentation results. In the field of supervised learning, Kittler *et al.*[33] summarized the different schemes for combining results from multiple classifiers. In the field of unsupervised clustering, Ghaemi *et al.*[34] performed a survey of methods in clustering ensembles. As far as the applications in medical imaging field, Grady[35] and Wattuya *et al.*[36] developed an algorithm to combine multiple segmentation results using the random walker method for natural image segmentations. Rohlfing *et al.*[37] studied atlas-based segmentation of biomedical images. They proposed to estimate the performances of the base classifiers and combine their respective

outputs by weighting them according to their estimated performance. This method is realized as a multiclass extension of an EM algorithm for ground truth estimation from a binary classification based on decisions of multiple experts.[38] Aljabar *et al.*[39,40] applied the majority voting rule[33] to combine segmentation results from atlas-based segmentation and presented a thorough evaluation on brain MR images. Ensemble segmentation showed its potential in these applications and we will apply it to the application of GBM brain tumors.

In this study, we propose an ensemble technique, applied to semiautomated GBM brain tumor segmentation on T1w postcontrast volumetric MR images, and evaluate the performance on a dataset with 46 tumor cases from a clinical trial research database. There are two steps involved. The first step is to generate input segmentation candidates from different algorithms. Three general-purpose segmentation methods were applied to generate input segmentations: fuzzy connectedness,[9] GrowCut,[41] and voxel classification using support vector machines (SVM).[42] The second step is to combine them to generate a final result. The ensemble scheme was confidence-based averaging (CMA). The CMA method was adopted based on an assumption that the majority of the base methods are correct, and errors from each method are independent so that they will be averaged out in the ensemble result. To our knowledge, we are the first to investigate ensemble segmentation for GBM tumor segmentation on single-channel MR images (T1w postcontrast), and to evaluate base methods and their ensemble on a relatively large dataset of 46 GBM tumors including different types of GBM tumor appearance patterns, in comparison to dataset of 5–20 cases in the prior literature.

2. MATERIALS AND METHODS

2.A. Input segmentations

We explored three algorithms as base methods including two semiautomated methods and one learning-based technique: fuzzy connectedness, GrowCut, and voxel classification using SVM. The fuzzy connectedness method was selected because it was reported to work well for semiautomated GBM brain tumor segmentation by Liu *et al.*[9] The GrowCut (GC) method was chosen due to its simple user interaction mechanism, straightforward implementation, and promising performance in our pilot study.[43] For these two semiautomated methods, user input seeds are provided in the tumor and background regions. SVM classification was chosen as a general-purpose method and adapted to this specific application by learning from examples.

2.A.1. Fuzzy connectedness

The fuzzy connectedness (FC) segmentation framework assigns fuzzy affinities to the target object during classification, to capture global "hanging togetherness" of voxels. The first step of the algorithm involves computing an "affinity" map, a local fuzzy relation, which quantifies the connectedness of any pixel pair in the original image; the second step calculates the "fuzzy connectedness," the global fuzzy relation with one specific (designated) pixel belonging to the object of interest.

We implemented the algorithm following Liu *et al.*'s work[9] since it has been previously applied to the GBM brain tumor segmentation task. The affinity between any two voxels c and d, denoted by $\mu_k(c, d)$, is given by

$$\mu_k = \begin{cases} 1 & \text{if } c = d, \\ 0 & \text{if } c \text{ and } d \text{ are not 6-adjacent,} \\ h_1(f(c), f(d)) * h_2(f(c), f(d)) & \text{otherwise,} \end{cases}$$

where $f(c)$ and $f(d)$ denote voxel intensity values at c and d, respectively.

The functional forms for h_1 and h_2 are chosen as follows:

$$h_1(f(c), f(d)) = \exp\left((-1/2)\left[\left(\left|\frac{f(c)-f(d)}{f(c)+f(d)}\right| - m_1\right) \Big/ s_1\right]\right),$$

$$h_2(f(c), f(d)) = \begin{cases} 0 & \text{if } f(c) + f(d) < a1, \\ \frac{f(c)+f(d)-a1}{a2-a1} & \text{if } a1 < f(c) + f(d) < a2, \\ 1 & \text{if } f(c) + f(d) > a2, \end{cases}$$

m_1 is set to the mean of the relative intensity differences $|f(c) - f(d)/(f(c) + f(d)|$ computed for all six-adjacent voxel pairs (c, d) within the region. s_1 is set to twice the standard deviation of this relative difference among the user input seeds. a_1 is set to [mean − twice the standard deviation of intensity sums $f(c) + f(d)$ of all six-adjacent voxel pairs (c, d) within the region]. a_2 is set to [mean + (twice the standard deviation of intensity sums $f(c) + f(d)$ of all six-adjacent voxel pairs (c, d) within the region]. The strength of the fuzzy connectedness is calculated by dynamic programming. There are numerous paths between any two given voxels c and d. In each possible path, the "strength of connectedness" is simply the smallest pairwise neighboring fuzzy affinity along this path. Among all possible paths, the one with the largest strength is the fuzzy connectedness of the two voxels c and d. The pool O of voxels with nonzero membership value in the fuzzy subset satisfies all of the following conditions: (1) all seed voxels are in O; (2) for any two voxels c and d in O, their strength of connectedness $S(c, d) > \theta$; and (3) for any voxels c in O and d not in O, $S(c, d) < \theta$.

2.A.2. GrowCut

The GC method[41] is based on cellular automata theory. Formally, a cellular automaton (CA) is a triple (S, N, δ), where S is the state set, N is the neighborhood, and $\delta: S^N -> S$ is the local transition function, where S^N indicates the states of the neighborhood cells at a given time, while S is the state of the central cell at the next time step. In the GC method, the cells correspond to image voxels, and the cell state $S = (C, l, \theta)$ for each voxel consists of the image feature vector C which is intensity in this study, the label l indicating the category to which the voxel belongs, and the strength θ in the continuous range [0, 1] indicating the confidence in the current labeling.

The GC method uses CA theory to interactively label the image volume using user supplied seeds. The user starts the segmentation by supplying seed points comprising both tumor and background voxels, the seeds' labels are set to the respective category labels, while their strength is set to 1. This sets the initial state of the cellular automaton. Strengths for unlabeled cells are set to 0. In each iteration t, each cell tries to "attack" the neighboring voxels by calculating the local intensity similarity; accordingly, the label map and the strength map are updated until convergence. The algorithm converges to a stable configuration, where cell states no longer change. The pseudo code for the GC algorithm is shown in Fig. 3, where $N(p)$ is the 26-neighbor system of a voxel p in 3D, and g is a monotonically decreasing function bounded within [0, 1]:

$$g(x) = 1 - \frac{x}{\max\|C\|_2}.$$

2.A.3. Voxel classification using SVM

The support vector machine (SVM)[42] is a supervised learning algorithm. It constructs a separating hyperplane in a multidimensional feature space that maximizes the margin between two classes. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the samples from the two classes. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier. Given a set of n labeled data points $(x_1, y_1), (x_2, y_2), \ldots, x_n, y_n$ where $y_i = \pm 1$, SVM searches for an optimal separating hyperplane $\langle w, x \rangle + b = 0$, where where $w \in R^n$, $x \in R^n$, and $b \in R$.

During the classifier training, voxels from manually contoured tumors are used as positive (tumor) examples, and an equal number of voxels sampled outside the tumor are used as negative (background) examples. For each training sample, a set of imaging features are calculated: intensity, gradient magnitude, first-order Gaussian derivatives (in three directions), second-order Gaussian derivatives (six in total), and the three eigenvalues of the Hessian matrix. These features are calculated at three different scales: 1, 2, and 4 pixels. In total, we have 42 features derived from images as the sum feature vector.

To apply the voxel classification to the test scan, the set of 42 features is calculated for each voxel and input to the trained classifier, and for each voxel a score that it belongs to a tumor is computed ranging from 0 to 1.

2.B. Combining input segmentations by confidence map averaging

A voxelized confidence map (CM) is generated for each base segmentation method. For the SVM method, the output score map was used as the CM. For the GC method, a strength map is generated by the algorithm, and we transform the strength map into a confidence map by linearly rescaling the foreground strength to [0.5, 1] and the background strength into [0, 0.5]. For the fuzzy connectedness method, the membership value is linearly rescaled to [0, 1] as the CM.

Code 1 Automata evolution rule

```
// For each cell...
for ∀p ∈ P
    // Copy previous state
    l_p^{t+1} = l_p^t;
    θ_p^{t+1} = θ_p^t;
    // neighbors try to attack current cell
    for ∀q ∈ N(p)
        if g(‖C_p − C_q‖_2) · θ_q^t > θ_p^t
            l_p^{t+1} = l_q^t
            θ_p^{t+1} = g(‖C_p − C_q‖_2) · θ_q^t
        end if
    end for
end for
```

FIG. 3. Pseudo code of the cellular automata evolution rule. Adapted from Ref. 41.

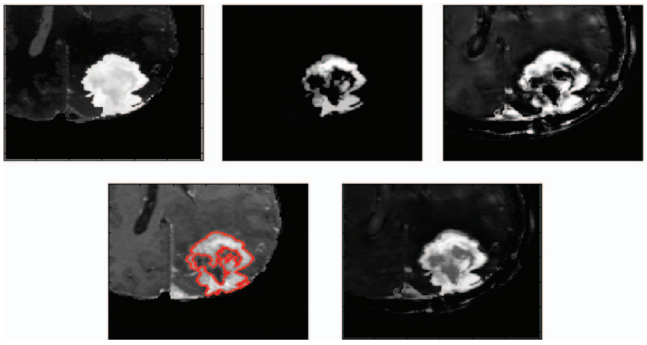

FIG. 4. Examples of confidence maps: the first row shows confidence maps of GC, FC, and SVM; the second row shows the original intensity image overlaid with the ground truth tumor contour and the averaged confidence map.

The three base methods ($N = 3$) are combined by confidence map averaging (CMA); the output of the ensemble is the average of the three confidence maps generated by the three base methods, weighting each of the three methods equally. Figure 4 shows the confidence maps from the three individual methods for one tumor. In order to obtain the binary segmentation, the CMA result is later thresholded to obtain a binary segmentation:

$$CMA(i, j, k) = \frac{1}{N}\sum_{n=1}^{N} CM_n(i, j, k).$$

3. EXPERIMENTS

We used 46 GBM tumor cases from 45 patients in this study from a 60-subject multisite research database. The 15 patients were excluded due to either lack of available manual gold standard tumor contours, anisotropy of voxel size, or variation in image resolution.

The imaging protocol for the T1w sequences was 3D volumetric acquisition in the axial plane using the flip angle-spoiled gradient echo sequence (FSPGR) or the magnetization-prepared rapid gradient-echo (MP-RAGE) sequence with 1 mm slice thickness, 0.9 mm by 0.9 mm pixel size, and 256*256 in-plane resolution.

The ground truth for the segmentation was manually contoured by a board-certified neuroradiologist with ten years of experience, with the facilitation of a semiautomated segmentation tool[44] from an in-house software system QIWS (quantitative imaging work station).

The brain volume was preprocessed to remove nonbrain matter and obtain consistent image intensities across all subjects for the given MR channel by the following steps: (1) skull-stripping — using FSL;[45] (2) B1 field correction and intensity normalization — using Freesurfer[46] to standardize the intensity of MR images acquired from different medical centers.

In order to reduce the processing time, we applied the algorithms in a predefined volume of interest (VOI). For each 3D MR volume, the user visually identified the start and end slice of the tumor, and provided manual seeds on the center slice of the tumor to initialize the GC method. With this information, the VOI can then be generated. First, the bounding box of the input seeds on the tumor center slice is extended 25 mm along each in-plane direction to enclose the whole tumor; then, the bounding box is extended in the z-direction to the start and end slice to obtain the VOI. Calculation time is thereby reduced by applying the segmentation framework only within the VOI instead of the whole brain volume.

We applied the proposed framework with the following parameter setup. For GC, users provided 3–5 seed points in the foreground and background structures respectively on the center slice of the tumor. The FC algorithm used the same foreground seeds, and the $s1$, $m1$, $a1$, and $a2$ were chosen as described in Sec. 2. The SVM voxel classification did not utilize the seeds. The SVM was trained for each leave-one-tumor-out iteration, resulting in 45 runs.

To obtain the binary segmentation results, the outputs of the SVM, fuzzy connectedness, and the CMA ensemble were thresholded adaptively using the Otsu method.[44] The binary segmentation from SVM and CMA were further processed by a connected component analysis to remove speckle noise involving: (1) removal of components smaller than 27 voxels

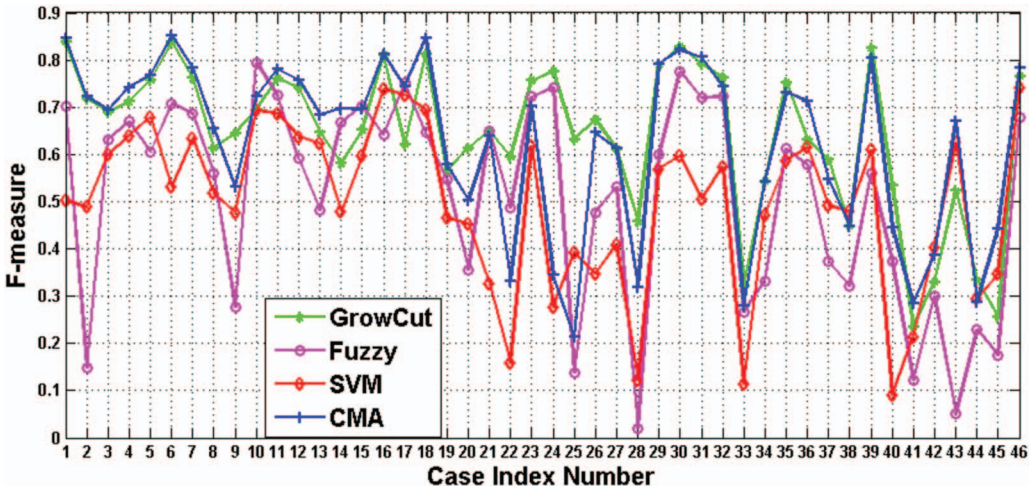

FIG. 5. F_1-measure of three base methods and the ensemble method for all 46 cases.

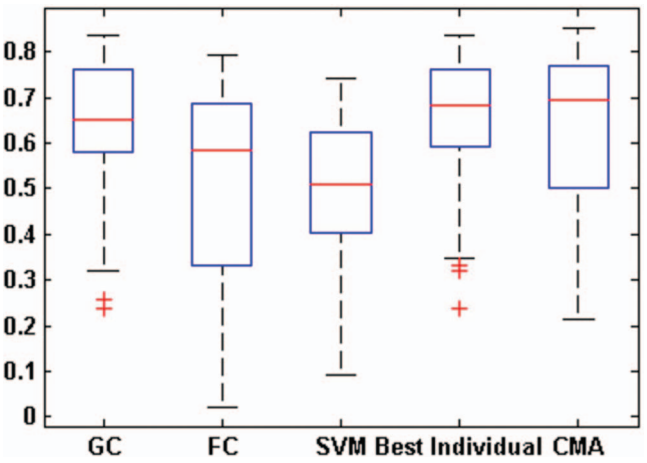

FIG. 6. Box plot of F_1-measure for three base methods (FC, GC, and SVM), the best individual result and the ensemble (CMA).

and (2) removal of components including background seeds but no foreground seeds.

The accuracy of the segmentation result was evaluated using the F_1-measure (ranging from 0 to 1)[47] between the semi-automated segmentation result and the ground truth.

$$F_1\text{-measure} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}},$$

$$\text{with} \quad \text{precision} = \frac{tp}{tp + fp}, \quad \text{recall} = \frac{tp}{tp + fn}.$$

4. RESULTS

We calculated the F_1-measure for all 46 GBM tumors to evaluate the accuracy of the segmentation results against the ground truth, and to compare the three base methods and our ensemble method. We present the F_1-measure plot for all 46 cases in Fig. 5.

First, to compare the three base methods, Fig. 5 shows that not a single base algorithm performs better than the other two algorithms in all the 46 cases. GC performed best for 34 cases out of 46, while FC and SVM performed best for seven and sic cases out of 46, respectively. The box plots and statistics of the the F_1-measures are shown in Fig. 6 and summary statistics are provided in Table I. A paired t-test was run to compare the three base methods the results are shown in Table II, indicating that GC and SVM are significantly better than FC method.

Second, to compare the ensemble with the three base methods, the ensemble method was close to the best base result for the majority of cases, although the best base method varied for each case. We obtained the best segmentation result for each case, and call it *best individual* result, and compared it with all other methods.The paired t-test shows that there is no significant difference between the best individual result and the ensemble result, while the best individual result is significantly better than all three base methods, as shown in Table III. The box plots and summary statistics of the *best individual* F_1-measures are shown in Fig. 6 and Table I.

TABLE I. Statistics of F_1-measure over 46 cases for different methods.

Different methods	Mean	Median	STD	IQR
FC	0.51	0.59	0.22	0.35
GrowCut	0.64	0.65	0.16	0.18
SVM	0.50	0.51	0.17	0.22
Best individual	0.66	0.68	0.15	0.17
CMA	0.63	0.7	0.18	0.27

TABLE II. Comparing different methods using paired t-test.

	FC	GrowCut	SVM	CMA
FC	N/A	$p < 0.05$	$p < 0.05$	$p < 0.05$
GrowCut	$p < 0.05$	N/A	$p < 0.05$	$p > 0.05$
SVM	$p < 0.05$	$p < 0.05$	N/A	$p < 0.05$
CMA	$p < 0.05$	$p > 0.05$	$p < 0.05$	N/A

The ensemble method improved the F_1-measure by approximately 0.04 (0.04 ± 0.02) compared to the *best individual* accuracy for 11 cases (no. 4, 8, 11, 12, 13, 14, 18, 36, 41, 43, 45), shown in Fig. 5. Two main reasons for the improvement are observed. One is that when the tumor is inhomogeneously enhanced, the ensemble method detected more tumor components than each base method. The other is that the necrosis was often falsely included as a part of a tumor by the GC and FC methods but correctly removed by the ensemble method. Figure 9 shows one example (index no. 12).

The ensemble method performs similar (0.0006 ± 0.01) to the *best individual* result for 21 cases (no. 1, 2, 3, 5, 6, 7, 15, 16, 17, 19, 21, 27, 29, 30, 31, 32, 34, 35, 39, 42, 46). Two main observations may contribute to this result. One observation is that one method (GC) performs relatively well when the tumor appears as a well-enhanced and single component, as shown in Fig. 10 with index no. 31, while the other two methods do not provide much additional value to the CMA method. The other observation is that in some cases the CMA not only includes more true positive voxels than the base methods, but also includes more false positive voxels, resulting in no overall improvement.

The ensemble method did not reach the performance of the *best individual* result (-0.13 ± 0.14) in 14 cases (no. 9, 10, 20, 22, 23, 24, 25, 26, 28, 33, 37, 38, 40, 44). Case no. 8 is an example shown in Fig. 11. In this case, the performance was reduced because of partial volumed voxels are missed by the ensemble method.

The ensemble method exhibited promising results in a subgroup of multifocal tumors. Multifocal tumors are those with more than one lesion site, as defined by intervening areas of normal brain signal, including or excluding the primary site,

TABLE III. Comparing the best individual segmentation result with others using paired t-test.

	FC	GrowCut	SVM	CMA
Best individual	$p < 0.05$	$p < 0.05$	$p < 0.05$	$p > 0.05$

all with a well-defined or mostly well-defined border. Figure 8 shows one example.[48] Cases 43–46 in Fig. 5 belong to this subgroup, and the zoom-in version is shown in Fig. 7 . For multifocal tumors, GC missed unconnected tumor pieces and SVM included all tumor pieces, and our ensemble method improved the performance over the GC method by 0.08 ± 0.01, improved over the SVM method by 0.04 ± 0.04, and improved over the FC method by 0.26 ± 0.25.

In general, the F_1-measure for all 46 cases is lower than 0.9 for all methods, because partial volumed voxels tended to be missed by the automated methods. Thus, the F_1-measure is reduced even when the segmentation result appears accurate by visual inspection.

5. DISCUSSION

In this study, we proposed an ensemble framework for the application of GBM brain tumor segmentation on high-resolution T1w postcontrast MR images. Rather than a highly customized method for this specific application, the proposed ensemble method can combine existing general-purpose segmentation algorithms to achieve greater consistency in performance.

Our study shows that ensemble segmentation has the potential to approximate the best result of the base method for each case. In spite of the power of the existing general-purpose segmentation methods, unfortunately not a single segmentation method could beat all the others in solving the challenging problem of GBM tumor segmentation with large variation of tumor appearance. We tested that ensemble segmentation has the potential to approximate the *best individual* result ($p > 0.05$), even though the *best individual* result is significantly better than all the base methods. In the future, with properly selected base methods which are good at segmenting different types of tumor appearances, an appropriate ensemble method may sustain the accuracy from the best "performer" for different tumor appearance and achieve an overall improvement over the base methods.

To our knowledge, we are the first to investigate and evaluate ensemble segmentation for GBM tumors on a relatively large dataset of 46 GBM tumors, in comparison to datasets of 5–20 cases in the literature. In the challenge of brain tumor segmentation (BRATS) at MICCAI2013, a training set of 30 patients were provided including both GBM and low grade gliomas. It is necessary to evaluate the GBM tumor segmentation over a large dataset including a variety of tumor presentations, because the appearance of GBM tumors on the images can vary substantially. GBM brain tumor segmentation is a challenging problem due to tumor heterogeneity, inhomogeneous intensity profiles, variable shapes and sizes, and recurrence patterns postsurgery. For example, there may or may not be necrosis/cavity/cyst present in the tumor core; the tumor recurrence may occur in the primary site or at a distant site; the tumor may show a vivid enhancement or diffuse pattern; and the tumor could have a blob shape or an irregular shape. Thus, it is crucial to evaluate the segmentation method on a large clinical dataset. Liu *et al.*[9] and Corso *et al.*[7] are until now the only two studies in the literature that evaluated their systems on a dataset of 20 cases. In our study, we included 46 tumors, including cases with all the clinical variability mentioned above. This study is thus significant in elucidating the range of tumor types to be addressed and thereby suggests that an ensemble approach may be appropriate.

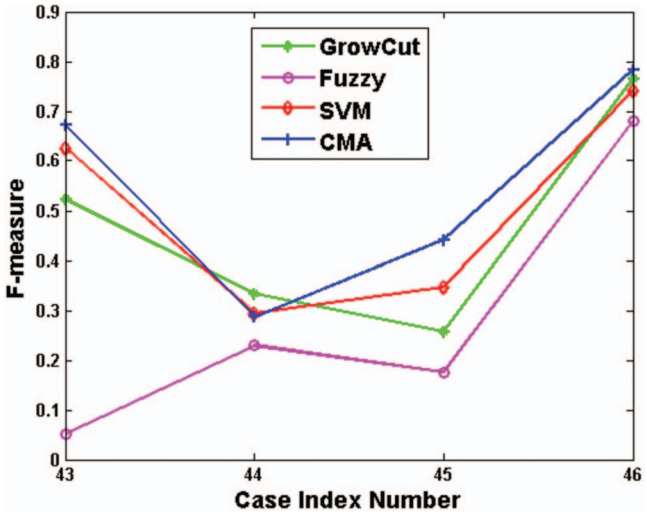

FIG. 7. F_1-measure of three base methods and the ensemble method for the subgroup of multifocal tumors.

To our knowledge, we are the first to investigate ensemble segmentation on single-channel MR images (T1w postcontrast) for GBM brain tumor segmentation. Most of previous studies developed fully automated segmentation using multichannel MR images (T1w, T2w, FLAIR, etc.), and we found only one publication, where Dube *et al.*[8] performed a preliminary study of fully automatic segmentation on this task using a dataset of seven patients. In the setting of GBM tumor clinical trials, radiologists manually contour contrast-enhanced

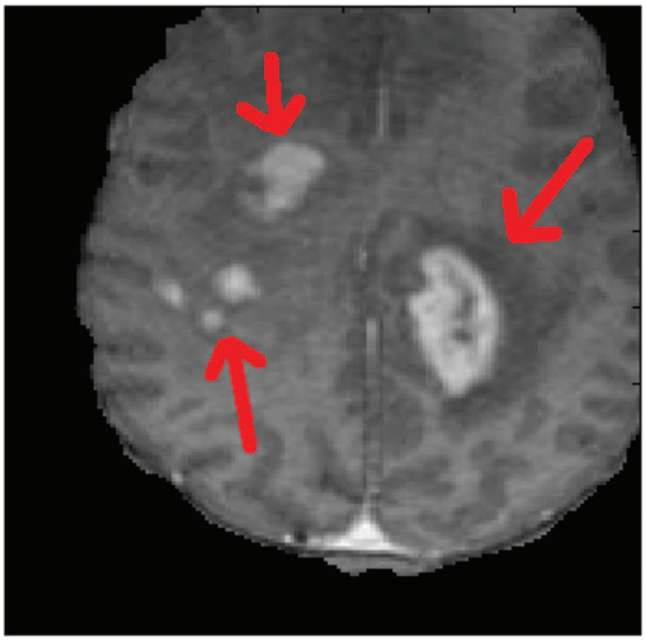

FIG. 8. An example of multifocal GBM tumor with arrows pointing to the contrast enhanced multifocal tumors.

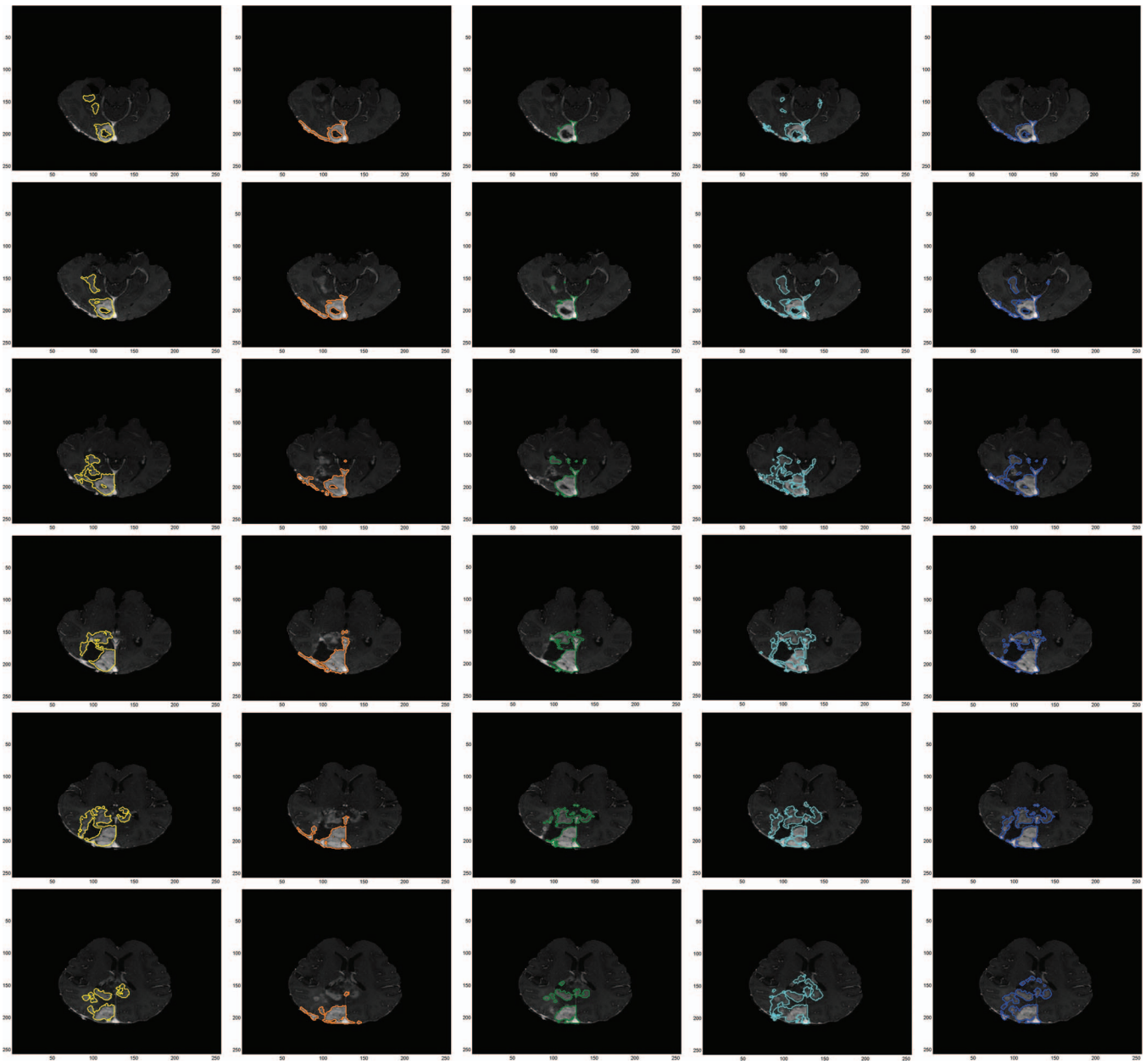

FIG. 9. Illustrative example of the segmentation results of the tumor with index number 12. Rows show results on different slices; columns show results using different segmentation methods: Column 1—the ground truth; Column 2—FC; Column 3—GC; Column 4—SVM; and Column 5—CMA ensemble.

tumors on single-channel T1w postcontrast images to measure tumor size change. Therefore, semiautomated segmentation on a single-channel T1w MR volume is relevant in a drug trial that uses radiographic response as a surrogate endpoint.

We compared the performance of different base segmentation algorithms on the application of GBM brain tumor segmentation. In the literature, many algorithms have been proposed as general-purpose segmentation methods; however, it is difficult to compare their performance since they were applied to different datasets. In this study, we evaluated three base algorithms on the same dataset, which serves as a reference to compare their performances and makes a useful contribution to the segmentation of GBM brain tumors.

Our study provides a potential general-purpose segmentation framework, even though our ensemble method was tested for a specific application of GBM tumor segmentation. In the context of tumor drug clinical trials where radiographical response is used as a surrogate endpoint, imaging core labs need a general-purpose segmentation method for medical image segmentation, and the ensemble framework is a potential solution. This is because imaging core labs collect and process data from different trials with different diseases and image modalities (CT, MRI, PET, etc.). It is tedious work for radiologists to manually contour the tumors, but it is expensive and inefficient to design a specific segmentation algorithm for each application. Therefore, a general purpose segmentation framework is attractive. However, medical image segmentation is not a trivial task due to the nature of medical image acquisitions and of heterogeneity of human diseases. An ensemble framework can take advantage of different

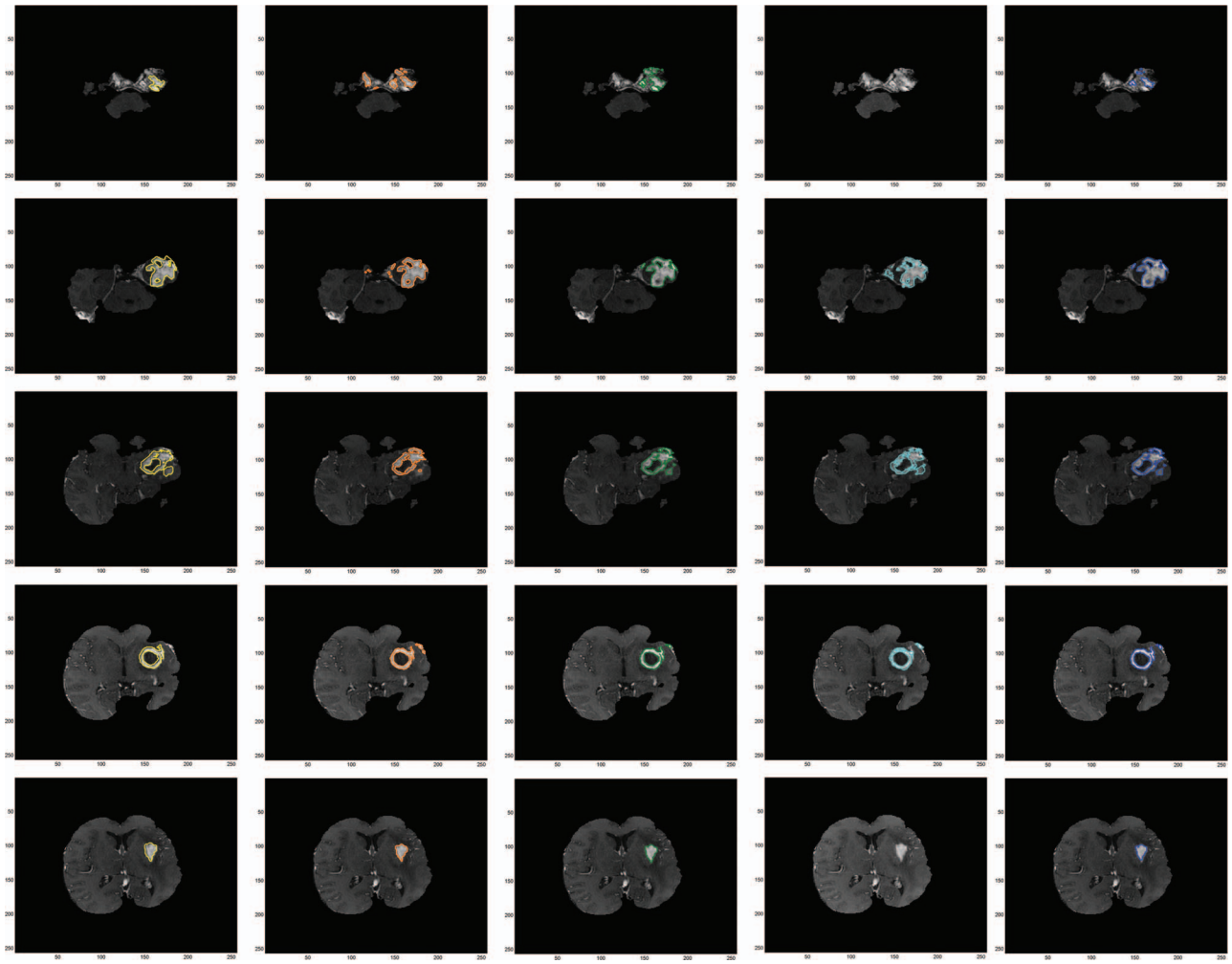

FIG. 10. Illustrative example of the segmentation results of the tumor with index number 31. Rows show results on different slices; columns show results using different segmentation methods: Column 1—the ground truth; Column 2—FC; Column 3—GC; Column 4—SVM; and Column 5—CMA ensemble.

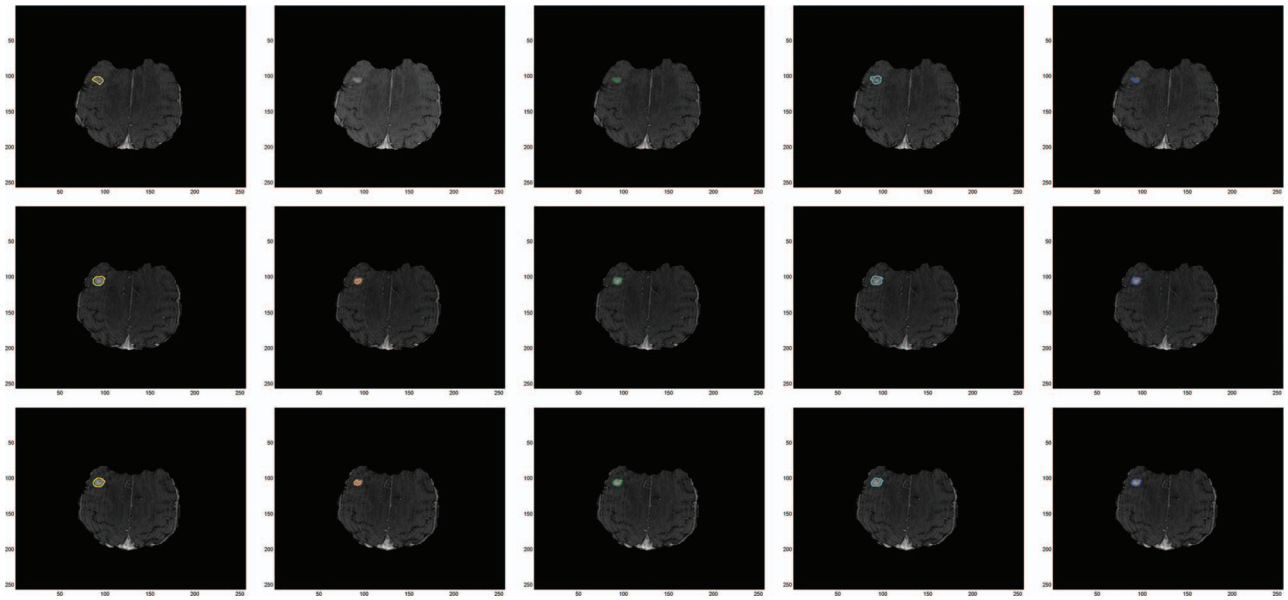

FIG. 11. Illustrative example of the segmentation results of the tumor with index number 8. Rows show results on different slices; columns show results using different segmentation methods: Column 1—the ground truth; Column 2—FC; Column 3—GC; Column 4—SVM; and Column 5—CMA ensemble.

segmentation algorithms. Its potential to serve as a general-purpose segmentation framework can be further studied and evaluated in other applications in the future to test the generalizability of the methods.

There are a couple of limitations in the present study design. One of them is the use of a single expert reading as the ground truth. Inter-observer variation is a limiting factor in GBM tumor segmentation due to the infiltrative nature and the boundary could be controversial sometimes, and it increases greatly in a post-operative setting, while the whole dataset in this study are post-operative scans. Meanwhile, segmentation for GBM tumor is still exploratory and in the current literature of GBM tumor segmentation, the majority of the studies are using a single reader for the evaluation. The reason is that, first, it is not easy to get large dataset with GBM tumors, and second, it is hard to collect results from multiple readers. Many of the literatures did not specify the ground truth, while in this paper, the ground truth is made by a neuroradiologist with ten-year experience in a real phase II drug trial. The other limitation is that the number of input seeds for FC and GC methods were not strictly controlled among all cases. As interactive method, GC is sensitive to user interactions. The amount of user interaction in this study is 5–9 seeds on the center slice of the tumor, to get an acceptable segmentation result. Thus, the segmentation result from GC is not the optimum from GC. The ensemble results might be even better if we get enough seeds to refine the GC results.

Future work could involve improvement of the CMA ensemble method. One possibility is to assign different weights to each base segmentation algorithm. Currently, we weighted all algorithms equally. Another possibility is to include additional base methods to explore whether more base methods can improve the segmentation performance.

In summary, we compared three base segmentation methods and evaluated the ensemble method on a clinical dataset of 46 GBM cases, and found that ensemble segmentation statistically approximates the *best individual* result ($p > 0.05$), and this provides motivation to investigate base methods that are good at segmenting tumors with different appearances. An ensemble method may then sustain the accuracy from the "best performer" for the various tumor appearances and obtain an overall improvement over the base methods.

[a]Author to whom correspondence should be addressed. Electronic mail: jinghuo@gmail.com

[1]D. Lipsitz, R. J. Higgins, G. D. Kortz, P. J. Dickinson, A. W. Bollen, D. K. Naydan, and R. A. Lecouteur, "Glioblastoma multiforme: Clinical findings, magnetic resonance imaging, and pathology in five dogs," Vet. Pathol. **40**, 659–669 (2003).

[2]E. G. Van Meir, C. G. Hadjipanayis, A. D. Norden, H. K. Shu, P. Y. Wen, and J. J. Olson, "Exciting new advances in neuro-oncology: The avenue to a cure for malignant glioma," Ca-Cancer J. Clin. **60**(3), 166–193 (2010).

[3]S. Drabycz, G. Roldán, P. de Robles, D. Adler, J. B. McIntyre, A. M. Magliocco, J. G. Cairncross, and J. R. Mitchell, "An analysis of image texture, tumor location, and mgmt promoter methylation in glioblastoma using magnetic resonance imaging," Neuroimage **49**, 1398–1405 (2010).

[4]W. E. Phillips, R. P. Velthuizen, S. Phupanich, L. O. Hall, L. P. Clarke, and M. L. Silbiger, "Applications of fuzzy c-means segmentation technique for tissue differentiation in MR images of a hemorrhagic glioblastoma multiforme," J. Magn. Reson. Imaging **13**, 277–290 (1995).

[5]M. C. Clark, L. O. Hall, D. B. Goldgof, R. Velthuizen, R. Murtagh, and M. S. Silbiger, "Automatic tumor segmentation using knowledge-based techniques," IEEE Trans. Med. Imaging **17**, 187–201 (1998).

[6]S. Ho, E. Bullitt, and G. Gerig, "Level set evolution with region competition: Automatic 3-d segmentation of brain tumors," in *Proceedings of the International Conference on Pattern Recognition*, Quebec, Canada (IEEE Computer Society, Quebec City, QC, Canada, 2002), pp. 532–535.

[7]J. J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille, "Efficient multilevel brain tumor segmentation with integrated Bayesian model classification," IEEE Trans. Med. Imaging **27**, 629–640 (2008).

[8]S. Dube, J. J. Corso, A. Yuille, T. F. Cloughesy, S. El-Saden, and U. Sinha, "Hierarchical segmentation of malignant gliomas via integrated contextual filter response," Proc. SPIE **6914**, 69143Y (2008).

[9]J. Liu, J. Udupa, D. Odhner, D. Hackney, and G. Moonis, "A system for brain tumor volume estimation via MR imaging and fuzzy connectedness," Comput. Med. Imaging Graph **29**, 21–34 (2005).

[10]L. M. Fletcher-Heath, L. O. Hall, D. B. Goldgof, and F. R. Murtagh, "Automatic segmentation of non-enhancing brain tumors in magnetic resonance images," Artif. Intell. Med. **21**, 43–63 (2001).

[11]M. Prastawa, E. Bullitt, S. Ho, and G. Gerig, "A brain tumor segmentation framework based on outlier detection," Med. Image Anal. **8**, 275–283 (2004).

[12]Z. Beevi and M. Sathik, "A robust segmentation approach for noisy medical images using fuzzy clustering with spatial probability," Int. Arab J. Inf. Technol. **9**(1), 74–83 (2012).

[13]H. Khotanloua, O. Colliotb, J. Atifc, and I. Bloch, "3d braintumor segmentationin MRI using fuzzy classification, symmetry analysis and spatially constrained deformable models," Fuzzy Sets Syst. **160**, 1457–1473 (2009).

[14]M. M. Ahmed and D. B. Mohamad, "Segmentation of brain MR images for tumor extraction by combining k-means clustering and Perona–Malik anisotropic diffusion model," Int. J. Image Process. **2**(1), 27–34 (2008).

[15]J. C. Bezdek, L. O. Hall, and L. P. Clarke, "Review of MR image segmentation techniques using pattern recognition," Med. Phys. **20**, 1033–1048 (1993).

[16]S. Vinitski, C. F. Gonzalez, R. Knobler, D. Andrews, T. Iwanaga, and M. Curtis, "Fast tissue segmentation based on a 4d feature map in characterization of intracranial lesions fast tissue segmentation based on a 4d feature map in characterization of intracranial lesions," J. Magn. Reson Imaging **9**, 768–776 (1999).

[17]S. K. Warfield, M. R. Kaus, F. A. Jolesz, and R. Kikinis, "Adaptive template moderated spatially varying statistical classification," Med. Image Anal. **4**(1), 43–55 (2000).

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

[19]M. Prastawa, E. Bullitt, N. Moon, K. V. Leemput, and G. Gerig, "Automatic brain tumor segmentation by subject specific modification of atlas priors," Acad. Radiol. **10**, 1341–1348 (2003).

[20]K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain," IEEE Trans. Med. Imaging **18**(10), 885–896 (1999).

[21]N. Moon, E. Bullitt, K. Van Leemput, and G. Gerig, "Automatic brain and tumor segmentation," *Proceedings of the Medical Image Computing and Computer-Assisted Intervention* (Springer, Tokyo, Japan, 2002), pp. 372–379.

[22]J. Zhang, K. Ma, M. H. Er, and V. Chong, "Tumor segmentation from magnetic resonance imaging by learning via one-class support vector machine," in *International Workshop on Advanced Image Technology* (IEEE, Singapore, 2004), p. 207–211.

[23]M. Schmidt, I. Levner, and R. Greiner, "Segmenting brain tumors using alignment-based features," in *Fourth International Conference on Machine Learning and Applications* (IEEE, Los Angeles, CA, 2005).

[24]C. H. Lee, M. Schmidt, A. Murtha, A. Bistritz, J. Sander, and R. Greiner, "Segmenting brain tumor with conditional random fields and support vector machines," in *Proceedings of Workshop on Computer Vision for Biomedical Image Applications at International Conference on Computer Vision* (Springer, Beijing, China, 2005).

[25]R. Ayachi and N. B. Amor, "Brain tumor segmentation using support vector machines," in *ECSQARU '09: Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Berlin, Heidelberg (Springer-Verlag, Germany, 2009), pp. 736–747.

[26]N. Zhang, S. Ruan, S. Lebonvallet, Q. Liao, and Y. Zhu, "Multi-kernel sSVM based classification for brain tumor segmentation of MRI multi-sequence," in *ICIP'09: Proceedings of the 16th IEEE International Conference on Image Processing*, Piscataway, NJ, USA (IEEE Press, Piscataway, NJ, 2009), pp. 3337–3340.
[27]K. Popuria, D. Cobzasb, M. Jagers, and S. L. Shaha, "3d variational brain tumor segmentation on a clustered feature set," in *Proceedings of SPIE Medical Imaging* (2009), Vol. 7258, pp. 72591N–72591N-10.
[28]S. Taheri, S. H. Ong, and V. F. H. Chong, "Level-set segmentation of brain tumors using a threshold-based speed function," Image Vision Comput. **28**(1), 26–37 (2010).
[29]J. J. Corso, A. L. Yuille, N. L. Sicotte, and A. W. Toga, "Detection and segmentation of pathological structures by the extended graph-shifts algorithm," in *Proceedings Medical Image Computing and Computer Aided Intervention (MICCAI)* (Springer, Australia, 2007), Vol. 1, pp. 985–994.
[30]S. Chandra, R. Bhat, and H. Singh, "A PSO based method for detection of brain tumors from MRI," in *Nature and Biologically Inspired Computing* (IEEE, India, 2009), pp. 666–671.
[31]Q. Ain, I. Mehmood, S. Naqi, and M. Jaffar, "Bayesian classification using DCT features for brain tumor detection," in *Knowledge-Based and Intelligent Information and Engineering Systems*, Lecture Notes in Computer Science Vol. 6276, edited by R. Setchi, I. Jordanov, R. Howlett, and L. Jain (Springer, Berlin/Heidelberg, 2010), pp. 340–349.
[32]A. Lashkari, "A neural network based method for brain abnormality detection in MR images using Gabor wavelets," Int. J. Comput. Appl. **4**(7), 1–8 (2010).
[33]J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998).
[34]R. Ghaemi, M. N. Sulaiman, H. Ibrahim, and N. Mustapha, "A survey: Clustering ensembles techniques," World Acad. Sci, Eng. Technol. **26**, 636–645 (2009).
[35]L. Grady, "Random walks for image segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1768–1783 (2006).
[36]P. Wattuya, K. Rothaus, J. S. Praßni, and X. Jiang, "A randomwalker based approach to combining multiple segmentations," *Proceedings of International Conference of Pattern Recognition (ICPR), Tampa, USA* (IEEE, Tampa, FL, 2008).
[37]T. Rohlfing, D. B. Russakoff, and C. R. Maurer, Jr., "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation," IEEE Trans. Med. Imaging **23**(8), 983–994 (2004).
[38]S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation," IEEE Trans. Med. Imag. **23**, 903–921 (2004).
[39]R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," Neuroimage **33**, 115–126 (2006).
[40]P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," Neuroimage **46**, 726–738 (2009).
[41]V. Vezhnevets and V. Konouchine, "Growcut: Interactive multi-label n-d image segmentation by cellular," *GraphiCon Proceedings* (Russia, 2005), pp. 150–156 [online].
[42]R. O. Duda, P. E. Hart, and D. H. Stork, *Pattern Classification* (Wiley Interscience, New York, 2000).
[43]J. Huo, E. M. van Rikxoort, K. Okada, H. J. Kim, W. Pope, J. Goldin, and M. Brown, "Confidence-based ensemble for gbm brain tumor segmentation," *Proceedings of the SPIE Medical Imaging* (SPIE, 2011), pp. 79622P–79622P-6.
[44]N. Otsu, "A threshold selection method from gray level histograms," IEEE Trans. Syst. Man Cybern. **9**, 62–66 (1979).
[45]S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. J. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J. M. Brady, and P. M. Matthews, "Advances in functional and structural MR image analysis and implementation as FSL," Neuroimage **23**, S208–S219 (2004).
[46]J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," IEEE Trans. Med. Imaging **17**, 87–97 (1998).
[47]C. J. van Rijsbergen, *Information Retrieval* (Butterworth, Newton, MA, 1979).
[48]W. B. Pope, Q. Xia, V. E. Paton, A. Das, J. Hambleton, H. J. Kim, J. Huo, M. S. Brown, J. Goldin, and T. Cloughesy, "Patterns of progression in patients with recurrent glioblastoma treated with bevacizumab," Neurology **76**(5), 432–437 (2011).