NeuroImage 47 (2009) 122–135

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Performance measure characterization for evaluating neuroimage segmentation algorithms

Herng-Hua Chang [a,*], Audrey H. Zhuang [b], Daniel J. Valentino [c,d], Woei-Chyn Chu [a]

[a] Institute of Biomedical Engineering, National Yang-Ming University, Taiwan
[b] Department of Radiation Oncology, City of Hope Cancer Center, USA
[c] Division of Interventional Neuro Radiology (DINR), UCLA, USA
[d] iCRco — The Innovative CR Company, CA, USA

ARTICLE INFO

Article history:
Received 18 September 2008
Revised 25 February 2009
Accepted 23 March 2009
Available online 5 April 2009

Keywords:
Segmentation
Evaluation
Accuracy
Precision
Jaccard
Dice
Tanimoto
Conformity
Sensitivity
Specificity
Sensibility

ABSTRACT

Characterizing the performance of segmentation algorithms in brain images has been a persistent challenge due to the complexity of neuroanatomical structures, the quality of imagery and the requirement of accurate segmentation. There has been much interest in using the Jaccard and Dice similarity coefficients associated with Sensitivity and Specificity for evaluating the performance of segmentation algorithms. This paper addresses the essential characteristics of the fundamental performance measure coefficients adopted in evaluation frameworks. While exploring the properties of the Jaccard, Dice and Specificity coefficients, we propose new measure coefficients Conformity and Sensibility for evaluating image segmentation techniques. It is indicated that Conformity is more sensitive and rigorous than Jaccard and Dice in that it has better discrimination capabilities in detecting small variations in segmented images. Comparing to Specificity, Sensibility provides consistent and reliable evaluation scores without the incorporation of image background properties. The merits of the proposed coefficients are illustrated by extracting neuroanatomical structures in a wide variety of brain images using various segmentation techniques.

Introduction

Image segmentation is a process of partitioning images into meaningful objects such that each region has similar characteristics reflected by their gray-level and texture. It is a critical pre-processing step in neuroimage analyses such as feature extraction, shape representation and measurement, and image understanding. With recent advances in the speed and resolution of medical imaging modalities and the increasing demand of brain imaging procedures, computerized assistance in facilitating the processing and analyses of image data has become important and necessary. The need for segmenting anatomical and pathological structures in brain images has increased dramatically in a number of clinic and research applications such as localization of pathology, quantitative volumetric assessment, surgical treatment planning, computer-aided diagnosis and surgery, brain mapping, vascular mapping, and 3-D visualization (Heinonen et al., 1999; Pham et al., 2000; Suri et al., 2002; Toga and Mazziotta, 2002; Ashburner and Friston, 2005; Joshi et al., 2007).

For several decades, a wide variety of human and animal brain image segmentation techniques have been proposed including threshold-based, region-based, statistics-based, classification, deformable models, atlas-guided techniques, and knowledge-based approaches (Frangi et al., 1999; Pham et al., 2000; Kaus et al., 2001; Dogdas et al., 2005; Sharief et al., 2008). To date, there is no universally acceptable segmentation technique that can produce satisfactory results in a broad range of neuroimage processing applications. Most algorithms make fundamental assumptions that limit their use for specific problems and applications (Dawant et al., 1999; MacDonald et al., 2000; Ali et al., 2005; Wu et al., 2006; Kloppel et al., 2008). While segmentation remains a challenging problem in brain image analyses, the development in the evaluation of segmentation algorithms has been lagging (Udupa et al., 2006).

Conceivably, this lag is the consequence of difficulty in defining the performance coefficients and statistics, difficulty in establishing gold standards, and exhaustibility in collecting tedious and time-consuming data (Haralick, 1994; Chalana and Kim, 1997; Udupa et al., 2006). Nevertheless, a number of evaluation methods have been proposed. Zhang (1996) suggested desirable properties of an evaluation method that include the generality for studying different segmentation algorithms, the ability for quantitative measurement, and the capacity

* Corresponding author.
E-mail address: emwave@ucla.edu (H.-H. Chang).

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.03.068

of evaluation on an objective basis. Objective study should exempt the influence of human factor and provide consistency and unbiased results. In addition, a good evaluation method should be able to detect small variations in segmented images.

The evaluation schemes are usually determined with respect to the goal of segmentation. Having appropriate coefficients for a particular application is essential to a successful evaluation. If the purpose of segmentation is to measure the volume of neuroanatomical structures, a volumetric error analysis may satisfy the needs. For analytical and systematical evaluation, more sophisticated approaches are required. Zhang (1996) divided the systematic evaluation methods into two categories: the analytical and empirical methods. The analytical methods directly examine and analyze the segmentation algorithms based upon their principles and properties, while the empirical ones adopt indirect judgments that apply the segmentation algorithms to test images and measure the quality of the results.

Recently, some researchers (Warfield et al., 2004; Fenster and Chiu, 2005; Udupa et al., 2006) suggested the report of accuracy, precision, and efficiency for characterizing the performance of image segmentation algorithms. Accuracy refers to the degree to which the segmentation results compare with the reference standards that represent the true segmentation. Precision refers to the repeatability of segmentation applying to the same image data. Efficiency provides information on the practical use of the algorithm, e.g., computational complexity and processing time.

The accuracy and precision evaluation can be broadly classified into distance-based coefficients, region-based coefficients, and statistical analyses of the entire images (Bland and Altman, 1986; Zhang, 1996; Chalana and Kim, 1997; Cox and Cox, 2000; Warfield et al., 2004; Fenster and Chiu, 2005). To accommodate various evaluation situations, multiple coefficients associated with statistical inferences, e.g., intra-class correlation, can be used. The distance-based coefficients [e.g., the Hausdorff distance (Huttenlocher et al., 1993)], based upon the measure of the distance between the segmentation contour (or surface in 3-D) and the true boundary, are used when the delineation of the boundary is critical. On the other hand, the region-based coefficients are used when the size and location measurement of the area (or volume) of the object is essential and is the objective of the segmentation. A variety of region-based similarity coefficients have been broadly studied by Cox and Cox (2000), some of which are summarized in Table 1.

Among a number of region-based coefficients based upon the measure of spatial overlap, the Jaccard (Jaccard, 1912) [alternatively known as the Tanimoto (Duda and Hart, 1973)] and Dice (Dice, 1945) coefficients have been extensively used for the performance evaluation of segmentation methods in brain images due to their simplicity (Vannier et al., 1991; Shan et al., 2002; Dogdas et al., 2005; Ashburner and Friston, 2005; Hernandez and Frangi, 2007; Joshi et al., 2007; Sharief et al., 2008). The Jaccard coefficient κ_j measures the ratio of the intersection area of two sets (Ω_1 and Ω_2) divided by the area of their union,

$$\kappa_j = \frac{|\Omega_1 \cap \Omega_2|}{|\Omega_1 \cup \Omega_2|} \times 100\%, \tag{1}$$

while the Dice coefficient κ_d, derived from a reliability measure known as the kappa statistic (Zijdenbos et al., 1994; Donner and Zou, 2002), computes the ratio of the intersection area divided by the mean sum of each individual area,

$$\kappa_d = \frac{2|\Omega_1 \cap \Omega_2|}{|\Omega_1| + |\Omega_2|} \times 100\%. \tag{2}$$

Table 1
Similarity coefficients for region-based evaluation.

Anderberg	$\frac{\theta_{TP}}{\theta_{TP} + 2(\theta_{FP} + \theta_{FN})}$
Blanque	$\frac{\theta_{TP}}{\max\{(\theta_{TP} + \theta_{FP}),(\theta_{TP} + \theta_{FN})\}}$
Dice	$\frac{2\theta_{TP}}{2\theta_{TP} + \theta_{FP} + \theta_{FN}}$
Jaccard	$\frac{\theta_{TP}}{\theta_{TP} + \theta_{FP} + \theta_{FN}}$
Kulczynski	$\frac{1}{2}\left(\frac{\theta_{TP}}{\theta_{TP} + \theta_{FP}} + \frac{\theta_{TP}}{\theta_{TP} + \theta_{FN}}\right)$
Ochiai	$\frac{\theta_{TP}}{[(\theta_{TP} + \theta_{FP})(\theta_{TP} + \theta_{FN})]^{\frac{1}{2}}}$
Simpson	$\frac{\theta_{TP}}{\min\{(\theta_{TP} + \theta_{FP}),(\theta_{TP} + \theta_{FN})\}}$

These two globally measured performance coefficients are often associated with the Sensitivity and Specificity coefficients that characterize how many pixels (or voxels) in the object are correctly segmented and how many pixels (or voxels) outside the object are correctly excluded, respectively.

Zijdenbos et al. (1994) used the Dice similarity coefficient for accessing the quantitation of white matter lesions in the human brain. Dawant et al. (1999) adopted the Dice coefficient to quantitatively evaluate their approach in the segmentation of internal structures of the head in magnetic resonance (MR) images. Shattuck et al. (2001) compared a partial volume tissue measurement model in classifying white matter and gray matter from nonbrain tissue by using the Jaccard and Dice coefficients. Shan et al. (2002) also used the Jaccard and Dice coefficients to evaluate their brain segmentation algorithm on 20 normal MRI datasets obtained from the Internet Brain Segmentation Repository (IBSR) (MGH, 2007). Recently, Crum et al. (2006) proposed the use of Jaccard incorporated with fuzzy set theory to evaluate segmentation and registration techniques. Udupa et al. (2006) proposed a framework for evaluating image segmentation algorithms by computing a minimum set of seven parameters that include the Jaccard, Sensitivity, and Specificity coefficients. Wu et al. (2006) and Kloppel et al. (2008) used the Sensitivity and Specificity coefficients to evaluate their automatic segmentation methods of MR images in multiple sclerosis lesion and Alzheimer's disease (AD), respectively. Powell et al. (2008) compared the automated segmentation algorithms for subcortical and cerebellar brain structures based upon the Jaccard, Dice, and Sensitivity coefficients.

This paper addresses the intrinsic properties and essential characteristics of the described performance measure coefficients, which have been widely adopted as fundamental elements in many evaluation frameworks. While exploring the properties of the Dice and Jaccard coefficients and their relationship, we also propose a new performance measure coefficient Conformity for characterizing the global error with respective to correct segmentation. Mathematically, the Conformity coefficient provides a much wider score range than the Jaccard and Dice coefficients for assessing segmentation results. In addition, we propose the Sensibility coefficient for measuring the error of background pixels (or voxels) being included without the incorporation of image dimensions as compared to the Specificity coefficient. The merits of these two new performance measure coefficients will be shown and demonstrated using a variety of segmentation techniques to extract anatomical and pathological structures in various brain image datasets.

The remainder of this paper is organized as follows. Methodology section introduces the definitions of the new coefficients Conformity and Sensibility. The relationship among the Conformity, Jaccard, and Dice coefficients is derived and established. We demonstrate typical scores of the Conformity and Sensibility coefficients using various overlapping scenarios. The intrinsic properties of the Jaccard, Dice, Conformity, Specificity, and Sensibility coefficients are also demonstrated. Results section illustrates the merits of the proposed coefficients by assessing a variety of segmentation results obtained from various algorithms on phantom and brain images in both 2-D and 3-D. Finally, in Discussion and conclusions section, the essential characteristics of different performance measure coefficients associated with the results are discussed and the contributions of this paper are summarized.

Methodology

Conformity

A quantitative performance analysis of segmentation results is important due to the fact that there is no perfect segmentation algorithm that can work for all types of brain images and usually have limited accuracy and precision. One way to assess the performance of algorithms is to measure the disparity between a segmented image by computerized methods and a reference image that is the best expected result (Zhang, 1996). This reference image, often called gold standard or ground truth, and the segmented image are obtained from the same input image being processed. For synthetic images, the true segmentation can be easily obtained from image generation procedures. However, when dealing with real images of patients or animal models, the truth is unknown and manually delineated images by experts are often used as references. According to the study by Fiez et al. (2000) on the intra- and inter-observer reliability, manual segmentation results produced high agreement within and between observers.

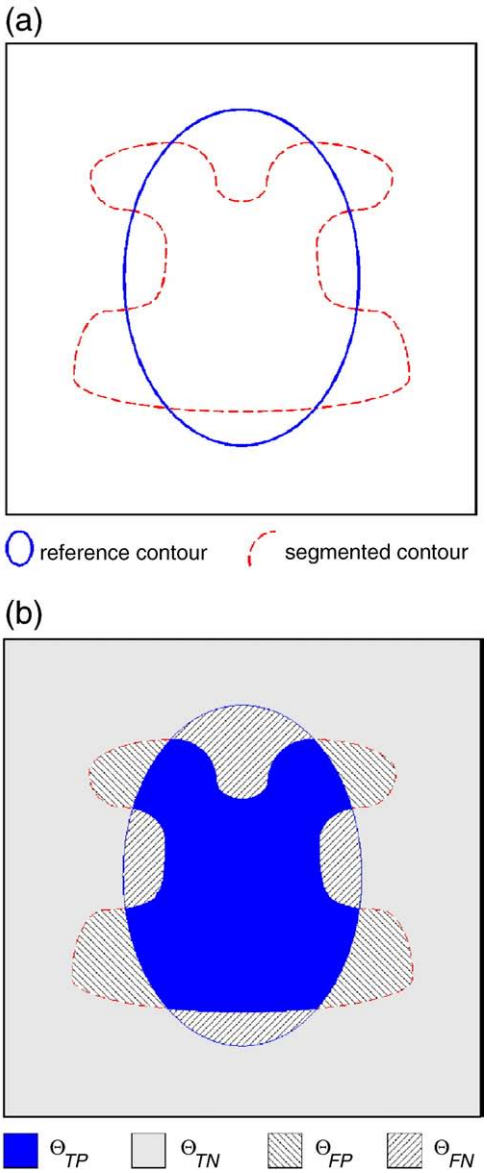

Fig. 1. Schematic illustrating the definition of fuzzy regions in computing performance measure coefficients. (a) Fuzzy segmentation results (the red dashed contour, Ω_{fs}) superimposed on the fuzzy reference standard (the blue solid contour, Ω_{fr}). (b) Definition of fuzzy regions representing Θ_{TP}, Θ_{TN}, Θ_{FP}, and Θ_{FN}.

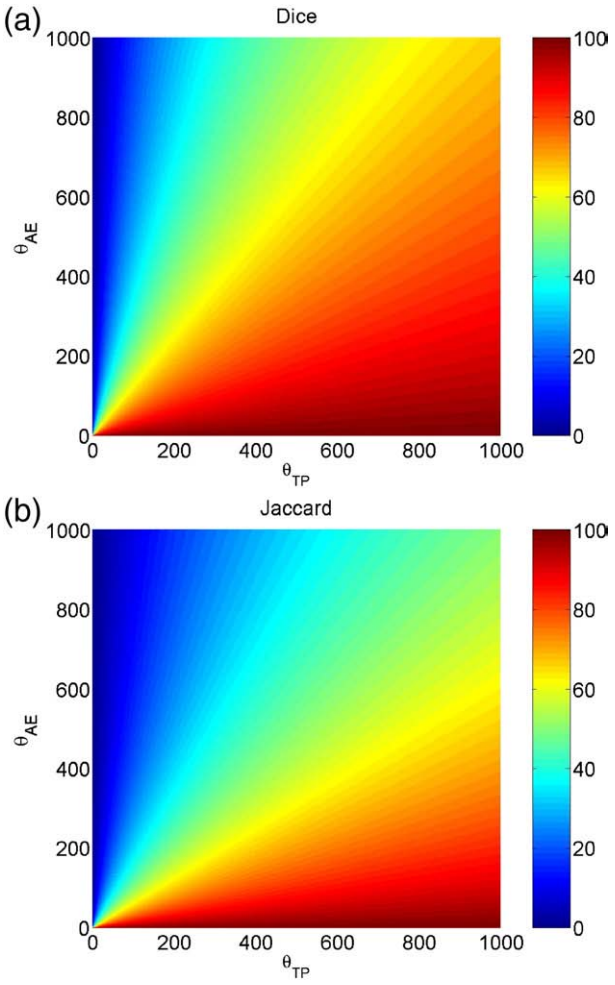

Fig. 2. Intrinsic properties of the Dice and Jaccard coefficients. (a) Score distribution of Dice with respect to Θ_{AE} and Θ_{TP}. (b) Score distribution of Jaccard with respect to Θ_{AE} and Θ_{TP}. The Dice coefficient is inherently inclined to produce higher scores than the Jaccard coefficient.

With recent advances in segmentation techniques using statistical and variational approaches, many algorithms provide a so-called "sub-pixel" (or "sub-voxel") estimate of the segmentation results. Rather than each pixel (or voxel) being labeled as belonging or not belonging to a region, i.e., binary labeling {0, 1}, there is a notation of partial association. To accommodate methods that provide fuzzy segmentation results, we denote the fuzzy segmentation as Ω_{fs} and the fuzzy reference Ω_{fr}, whose partial association taking values in the continuous domain [0, 1]. The fuzzy reference can be generated by averaging multiple manual delineations, which are inherently binary though. An alternative is to use the STAPLE algorithm (Warfield et al., 2004) for estimating the surrogate of true segmentation from a set of delineations produced by a group of experts. Similarly, the binary segmentation results can be converted into fuzzy segmentation by averaging the outcomes of multiple repetitions.

For the sake of simplicity, let us first consider the binary case of fuzzy masks obtained from the reference (the blue solid contour) and a segmentation method (the red dashed contour) as depicted in Fig. 1a. A global similarity coefficient called Conformity κ_c is

defined for measuring the ratio of the number of mis-segmented pixels (or voxels) to the number of correctly segmented pixels (or voxels) using the following equation:

$$k_c \equiv \begin{cases} \left(1 - \frac{\Theta_{FP} + \Theta_{FN}}{\Theta_{TP}}\right) \times 100\% = \left(1 - \frac{\Theta_{AE}}{\Theta_{TP}}\right) \times 100\% & \text{if } \Theta_{TP} > 0, \\ \text{Failure} & \text{if } \Theta_{TP} = 0, \end{cases} \tag{3}$$

where Θ_{FP} represents false positives of the fuzzy union, Θ_{FN} false negatives, and Θ_{TP} true positives as illustrated in Fig. 1b. In the above equation, $\Theta_{AE} = \Theta_{FP} + \Theta_{FN}$ represents all errors of the fuzzy segmentation results. A detail description of the computation of Θ_{TP}, Θ_{TN}, Θ_{FP}, and Θ_{FN} based upon the fuzzy regions is provided in the Appendix. Mathematically, κ_c can be negative infinity if $\Theta_{TP} = 0$. This is the case when the fuzzy segmentation Ω_{fs} and the fuzzy reference Ω_{fr} are totally apart without any overlap. Such a segmentation result is definitely inadequate and treated as *failure* without the need of any further analysis as shown in Eq. (3).

Having the performance measure coefficients expressed in terms of individual fuzzy regions (e.g., Θ_{AE} and Θ_{TP}) enables better interpretation and understanding of intrinsic properties. We do this by rewriting the Jaccard coefficient κ_j in Eq. (1) and the Dice coefficient κ_d in Eq. (2) respectively as

$$k_j = \frac{\Theta_{TP}}{\Theta_{TP} + \Theta_{FP} + \Theta_{FN}} \times 100\% = \frac{1}{1 + \frac{\Theta_{AE}}{\Theta_{TP}}} \times 100\%, \tag{4}$$

$$k_d = \frac{2\Theta_{TP}}{2\Theta_{TP} + \Theta_{FP} + \Theta_{FN}} \times 100\% = \frac{2}{2 + \frac{\Theta_{AE}}{\Theta_{TP}}} \times 100\%. \tag{5}$$

Now, the κ_j and κ_d are expressed in terms of Θ_{AE} and Θ_{TP}, whose score distributions are shown in Fig. 2. It is observed that the scores of the Dice coefficient are more inclined towards high values while the Jaccard coefficient is more uniformly distributed.

To investigate the relationship among the Dice, Jaccard, and Conformity coefficients, we further define the discrepancy-to-concordance ratio ξ as

$$\xi = \frac{\Theta_{AE}}{\Theta_{TP}} \quad \text{if} \quad \Theta_{TP} \neq 0. \tag{6}$$

We can then express the Conformity, Jaccard, and Dice coefficients in terms of ξ by substituting Eq. (6) into Eqs. (3), (4), and (5) as

$$\kappa_c = 1 - \xi, \tag{7a}$$

$$\kappa_j = \frac{1}{1 + \xi}, \tag{7b}$$

$$\kappa_d = \frac{2}{2 + \xi}. \tag{7c}$$

The above equations indicate that the Jaccard and Dice coefficients nonlinearly respond to the discrepancy-to-concordance ratio ξ while the Conformity coefficient has a linear correspondence. This can be easily interpreted by plotting these performance measure coefficients with respect to ξ based upon Eq. (7a) to (7c) as depicted in Fig. 3a. The red dashed curve represents κ_d, the blue solid curve κ_j, and the black dotted line κ_c. It indicates that both Jaccard and Dice coefficients are bounded between 0% and 100%. The curves of κ_j and κ_d are nearly parallel with each other and approaching a horizontal line when ξ is larger. On the other hand, the Conformity coefficient linearly responds to the change of ξ. For example, $\kappa_c = -1000\%$ corresponds to $\xi = 11$, i.e., the number of incorrectly segmented pixels (or voxels) is 11 times larger than the number of correctly segmented pixels (or voxels).

One important property of a good evaluation coefficient is the discrimination capability to distinguish small variations in segmented images as described in Introduction. The discrimination capabilities of the Conformity, Jaccard, and Dice coefficients can be obtained by taking the derivatives of Eq. (7a) to (7c) with respect to ξ as

$$\kappa_c' = \frac{d\kappa_c}{d\xi} = -1, \tag{8a}$$

$$\kappa_j' = \frac{d\kappa_j}{d\xi} = \frac{-1}{(1 + \xi)^2}, \tag{8b}$$

$$\kappa_d' = \frac{d\kappa_d}{d\xi} = \frac{-2}{(2 + \xi)^2}. \tag{8c}$$

Fig. 3b shows the plots of the absolute values of Eq. (8a) to (8c) in the domain of $\xi = 0$ to 11. While Conformity has a constant value of unity regardless of ξ, the discrimination capabilities of Jaccard and Dice decrease rapidly and their values approach zero when ξ is larger. It is interesting to note that Jaccard and Dice both have the same discrimination value of 0.1716 when $\xi = \sqrt{2}$.

The discrimination values of Jaccard are higher than Dice when $\xi < \sqrt{2}$ and lower when $\xi > \sqrt{2}$.

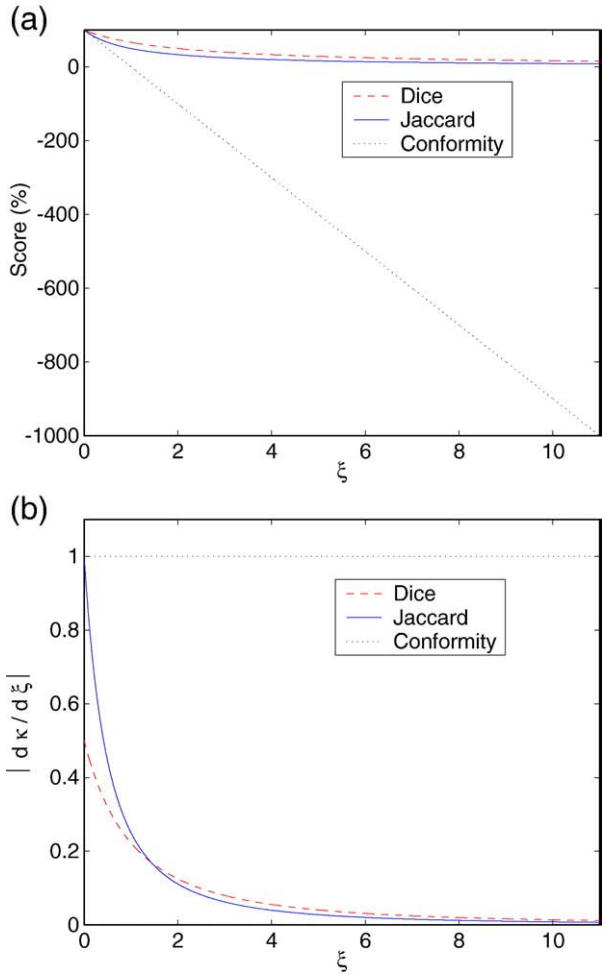

Fig. 3. Plots of the Dice (the red dashed curve), Jaccard (the blue solid curve), and Conformity (the black dotted line) coefficients with respect to the discrepancy-to-concordance ratio ξ. (a) Both Dice and Jaccard coefficients are bounded between 0% and 100%, while the Conformity coefficient linearly reflects the change of ξ. (b) The absolute values of the derivatives of the performance measure coefficients in (a) with respect to ξ.

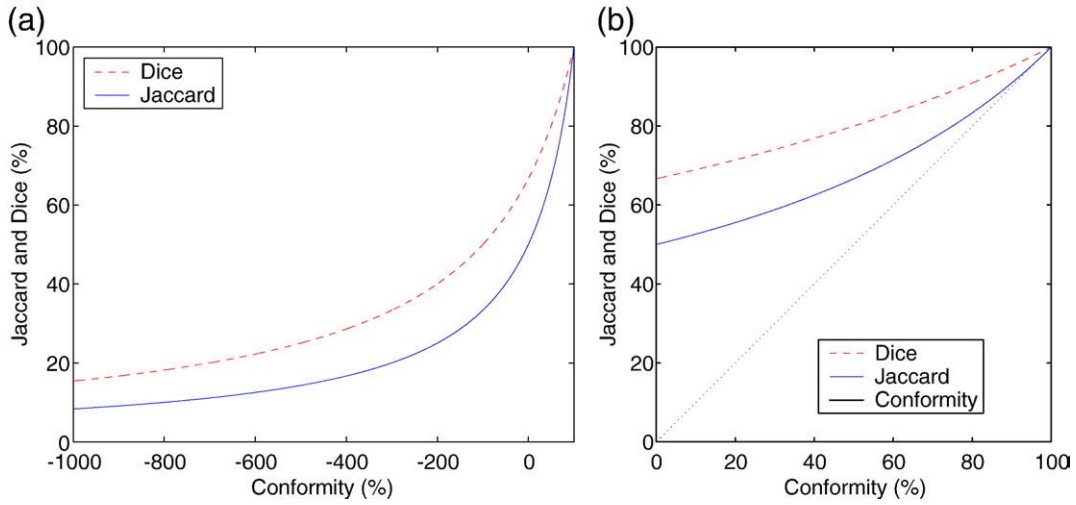

Fig. 4. Comparison of the Jaccard κ_j (the blue solid curve) and Dice κ_d (the red dashed curve) coefficients with respect to Conformity κ_c in the range of (a) −1000% to 100% and (b) 0% to 100%. The Conformity coefficient, which provides a much wider range of index scores, is always smaller than the other two coefficients κ_j and κ_d (except at 100%). For reference, $\kappa_c = 0\%$ corresponds to $\kappa_j = 50\%$ and $\kappa_d = 66.67\%$, respectively.

To more thoroughly study the essential characteristics of the Dice, Jaccard, and Conformity coefficients, we substitute Eq. (7b) into Eq. (7c) and relate the Jaccard and Dice coefficients as

$$\kappa_j = \frac{\kappa_d}{2 - \kappa_d}, \tag{9a}$$

$$\kappa_d = \frac{2\kappa_j}{\kappa_j + 1}. \tag{9b}$$

The Jaccard and Dice coefficients can be expressed in terms of Conformity by substituting Eq. (7a) into Eqs. (7b) and (7c) as

$$\kappa_j = \frac{1}{2 - \kappa_c}, \tag{10}$$

$$\kappa_d = \frac{2}{3 - \kappa_c}. \tag{11}$$

Finally, by rearranging Eqs. (10) and (11), the Conformity coefficient κ_c can be expressed in terms of κ_j and κ_d as

$$\kappa_c = \frac{2\kappa_j - 1}{\kappa_j}, \tag{12}$$

$$\kappa_c = \frac{3\kappa_d - 2}{\kappa_d}. \tag{13}$$

In Fig. 4, we show the plots of the Dice and Jaccard coefficients with respect to the Conformity coefficient in the domain from −1000% to 100% based upon Eqs. (10) and (11). The three coefficients agree with one another when the two fuzzy sets (Ω_{fs} and Ω_{fr}) are perfectly matched with a uniform score of 100%. Both κ_j and κ_d have a minimum score of zero when there is no intersection at all, i.e., $\Theta_{TP} = 0$ [see Eqs. (4) and (5)]. The Conformity coefficient κ_c is always smaller than the other two coefficients and has a much wider range of index scores $(-\infty, 1]$. For reference, $\kappa_c = 0\%$ corresponds to $\kappa_j = 50\%$ and $\kappa_d = 66.67\%$, respectively. A zero score of κ_c is obtained when the number of correctly segmented pixels (or voxels) equals to the number of mis-segmented pixels (or voxels), i.e., $\Theta_{TP} = \Theta_{AE}$, and negative scores are obtained when $\Theta_{TP} < \Theta_{AE}$.

Sensibility

While the global performance measure coefficients provide distinct assessment on how the fuzzy segmentation matches the fuzzy reference as described in Conformity section, researchers are also interested in realizing how much the region of interest is being excluded and included, respectively. In addition to the global measure coefficients, the Sensitivity and Specificity coefficients are often used to complement the evaluation of segmentation algorithms (Shattuck et al., 2001; Fenster and Chiu, 2005; Udupa et al., 2006; Wu et al., 2006; Kloppel et al., 2008; Powell et al., 2008). The Sensitivity coefficient η_{stv} is used

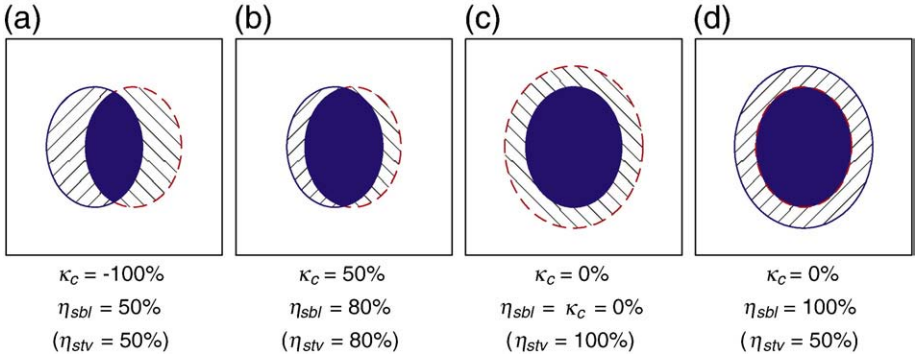

Fig. 5. Schematic illustrating typical scores of Conformity and Sensibility with various overlapping scenarios. (a) Half-overlap with $\Theta_{TP} = \Theta_{FP} = \Theta_{FN}$, i.e., $\Theta_{AE} = 2\Theta_{TP}$. (b) Close-overlap with $\Theta_{TP} = 4\Theta_{FP} = 4\Theta_{FN}$, i.e., $\Theta_{TP} = 2\Theta_{AE}$. (c) Encompassed-overlap with $\Theta_{TP} = \Theta_{FP} = \Theta_{AE}$ and $\Theta_{FN} = 0$. (d) Interior-overlap with $\Theta_{TP} = \Theta_{FN} = \Theta_{AE}$ and $\Theta_{FP} = 0$.

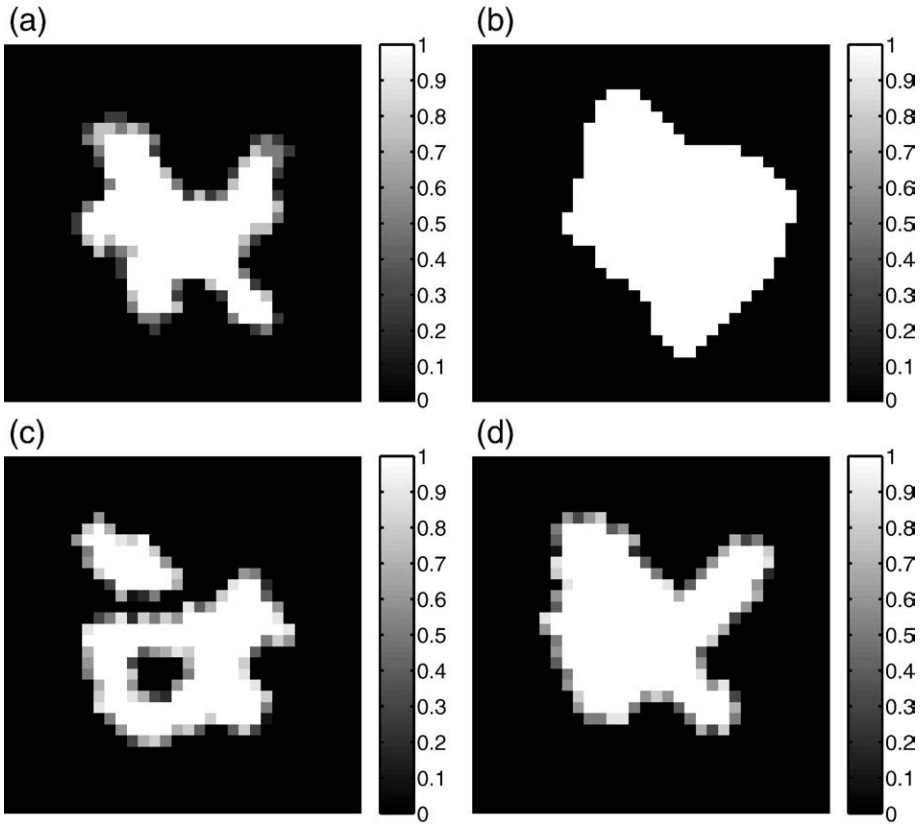

Fig. 6. Computation of the performance measure coefficients based upon 32×32 fuzzy masks. (a) The mask of the fuzzy reference. (b) The mask obtained using binary segmentation. (c) The mask obtained using region-based segmentation. (d) The mask obtained using contour-based segmentation. Note that all masks were with fuzzy representations except (b), which was binary.

for measuring how many pixels (or voxels) in Ω_{fr} are correctly segmented using

$$\eta_{stv} = \left(\frac{\theta_{TP}}{\theta_{TP} + \theta_{FN}}\right) \times 100\% = \left(\frac{\theta_{TP}}{\Omega_{fr}}\right) \times 100\%, \tag{14}$$

while the Specificity coefficient η_{spf} measures how many pixels (or voxels) outside Ω_{fr} are correctly excluded using

$$\eta_{spf} = \left(\frac{\theta_{TN}}{\theta_{TN} + \theta_{FP}}\right) \times 100\%. \tag{15}$$

Note that the Specificity coefficient involves the computation of true negatives Θ_{TN} of the fuzzy union, which is related to the dimension of images. In Results section, the inconsistency of using Specificity for evaluation due to Θ_{TN} will be demonstrated.

To faithfully reflect the amount of disagreement and not to depend on the dimension of images associated with Θ_{TN}, a new coefficient

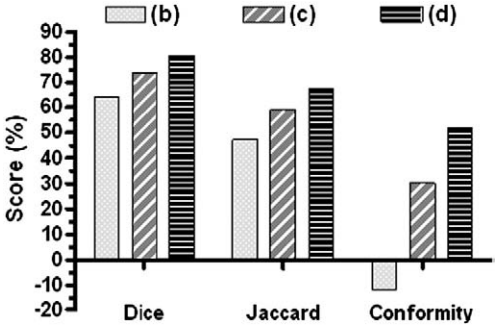

Fig. 7. Global evaluation for the results in Figs. 6b–d using the Dice, Jaccard, and Conformity coefficients, whose score ranges were 16.5%, 20.4%, and 63.8%, respectively.

Sensibility η_{sbl} is defined for measuring how many pixels (or voxels) outside Ω_{fr} are included as given in the following equation,

$$\eta_{sbl} \equiv \left(1 - \frac{\theta_{FP}}{\theta_{TP} + \theta_{FN}}\right) \times 100\% = \left(1 - \frac{\theta_{FP}}{\Omega_{fr}}\right) \times 100\%. \tag{16}$$

Mathematically, the values of η_{sbl} are negative if $\Theta_{FP} > \Theta_{TP} + \Theta_{FN} = \Omega_{fr}$. This is the case when the mis-segmented region outside Ω_{fr} is larger than the fuzzy reference region.

In Fig. 5, we illustrate typical scores for the Conformity and Sensibility coefficients using various overlapping scenarios (assuming binary cases). Scores of $\kappa_c = -100\%$ and $\eta_{sbl} = 50\%$ are obtained when Ω_{fs} and Ω_{fr} are half-overlap with $\Theta_{TP} = \Theta_{FP} = \Theta_{FN}$, i.e., $\Theta_{AE} = 2\Theta_{TP}$, as depicted in Fig. 5a. A scenario of $\kappa_c = 50\%$ with $\eta_{sbl} = 80\%$ is achieved when $\Theta_{TP} = 4\Theta_{FP} = 4\Theta_{FN}$, i.e., $\Theta_{TP} = 2\Theta_{AE}$, as illustrated in Fig. 5b. When Ω_{fs} entirely encompasses Ω_{fr} with $\Theta_{TP} = \Theta_{FP} = \Theta_{AE}$ and $\Theta_{FN} = 0$ as shown in Fig. 5c, one can obtain $\eta_{stv} = 100\%$ and $\eta_{sbl} = \kappa_c = 0\%$. When Ω_{fs} is entirely inside Ω_{fr} with $\Theta_{TP} = \Theta_{FN} = \Theta_{AE}$ and $\Theta_{FP} = 0$ as depicted in Fig. 5d, one can obtain $\kappa_c = 0\%$, $\eta_{sbl} = 100\%$, and $\eta_{stv} = \kappa_j$. Lastly, it is noted that Conformity is sensitive to both differences in dimension and location, though, differences in location are more distinctly reflected than differences in size.

Table 2
Evaluation results in Figs. 6b–d based upon the Sensitivity, Specificity, and Sensibility coefficients, whose score ranges were 14.8%, 9.7%, and 44.3%, respectively.

Image	Sensitivity	Specificity	Sensibility
Fig. 6b	82.29%	83.32%	25.70%
Fig. 6c	76.56%	93.02%	70.01%
Fig. 6d	91.41%	91.86%	64.72%

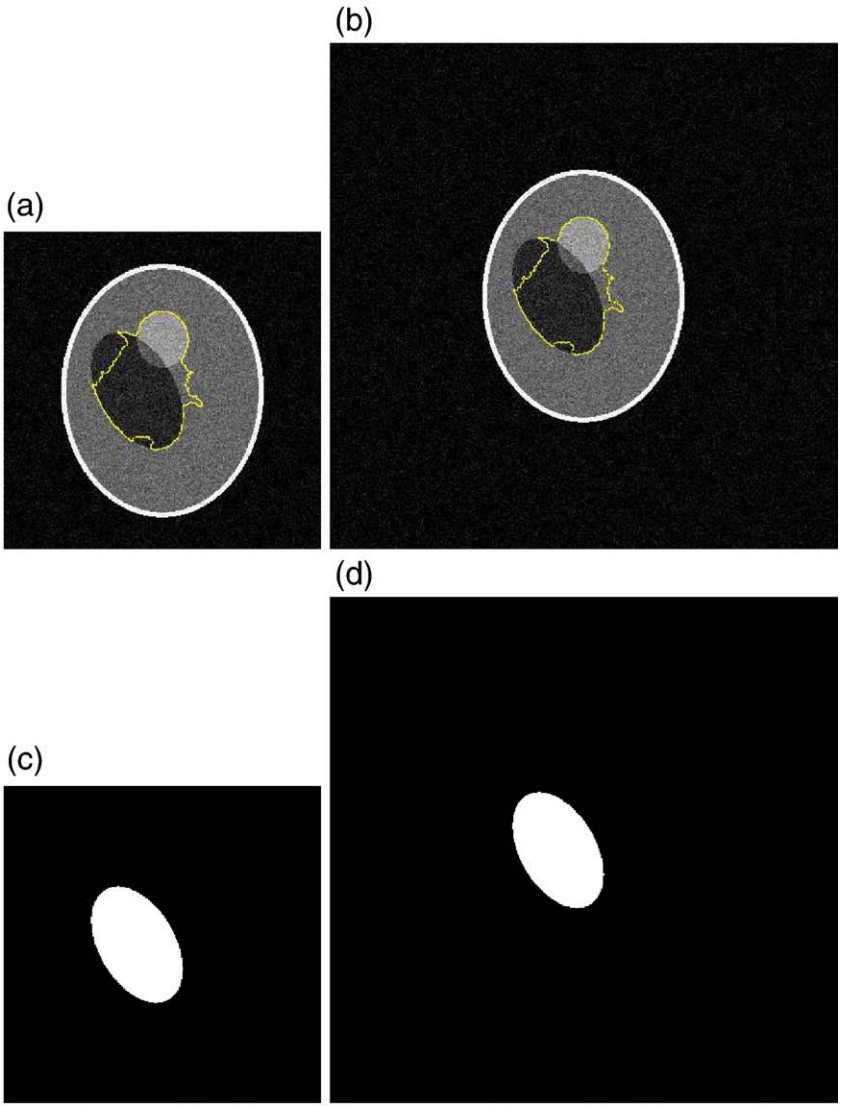

Fig. 8. Segmentation of a simulated neuroanatomical structure in phantom images with different dimensions corrupted by 10% Gaussian noise using the CFM algorithm. Note that the results in (a) 320×320 and (b) 512×512 were exactly identical and the corresponding reference masks are shown in (c) and (d), respectively. The scores of Specificity were different with 96.99% for (a) and 98.88% for (b), while the scores of Sensibility were consistent with a uniform value of 62.87% as presented in Table 3.

Image data and segmentation methods

A variety of image data are used to demonstrate the characteristics of the performance measure coefficients including both synthetic and real neuroimages. Simulated images of neuroanatomical structures were generated using the "phantom" command in the MATLAB® environment. T1-weighted simulated brain MR images were obtained from the BrainWeb simulator repository (Kwan et al., 1999) that had 9% noise and 40% intensity non-uniformity. The dimension of these images was 181×217 with spatial resolution 1×1 mm. Images with disease were acquired from the CCB Test Data Archive in the Laboratory of Neuro Imaging (LONI) and from the medical image database in the Division of Interventional Neuro Radiology (DINR) at UCLA under an approved Human Research Subject Protection Protocol. Finally, 20 normal brain MRI datasets obtained from the IBSR (MGH, 2007) were used to provide statistical analyses. Each of the image volumes had approximately 60 slices with 256×256 pixels per slice. The slice resolution varied from 1.02×1.04 to 1.26×1.51 mm and the inter-slice spacing varied from 3.05 to 3.37 mm.

Seven different segmentation algorithms were evaluated in extracting neuroanatomical structures based upon the described performance measure coefficients. Namely, they were the charged fluid model (CFM) (Chang, 2006), active contour model (ACM) (Kass et al., 1988), Chan–Vese level set (CVL) (Chan and Vese, 2001), Caselles–Malladi level set (CML) (Caselles et al., 1993; Malladi et al., 1995), brain extraction tool (BET) (Smith, 2002), brain surface extractor (BSE) (Shattuck et al., 2001), and model-based level set (MLS) (Zhuang et al., 2006) methods.

Results

We start by illustrating the computation of the performance measure coefficients in fuzzy cases and the comparison of different segmentation results based upon these measurements. As Fig. 6a shows, the fuzzy reference with values in the continuous domain $[0,1]$

Table 3
Evaluation analyses using a number of performance measure coefficients for the results in Fig. 8.

Size	Image	κ_d	κ_j	κ_c	η_{stv}	η_{sbl}	η_{spf}
320×320	Fig. 8a	76.78%	62.31%	39.52%	85.45%	62.87%	96.99%
512×512	Fig. 8b	76.78%	62.31%	39.52%	85.45%	62.87%	98.88%

Note that all coefficients had unanimous scores for both identical results, except Specificity η_{spf} that produced different scores for the same segmentation result.

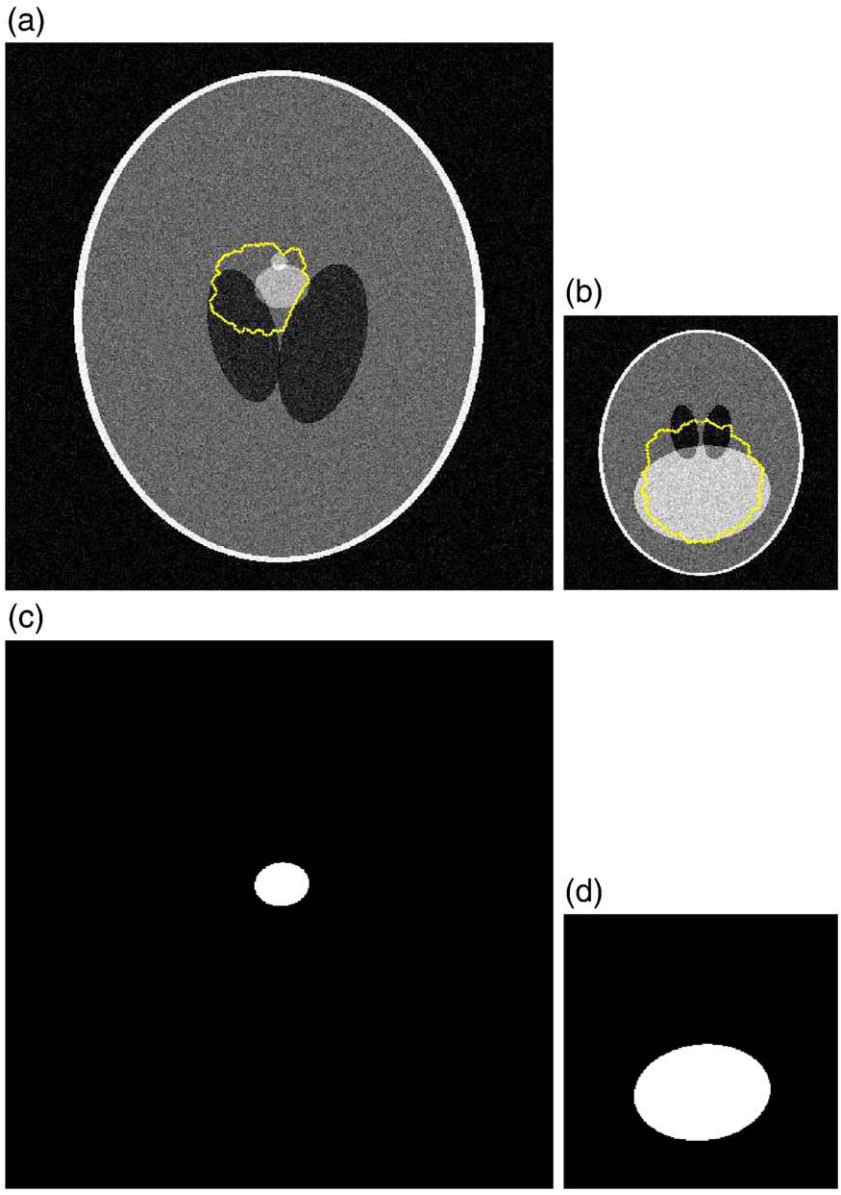

Fig. 9. Extraction of simulated neuroanatomical structures in noisy images with different dimensions using the ACM method. The dimensions of the segmented images were 512×512 for (a) and 256×256 for (b). The corresponding reference masks are shown in (c) and (d), respectively. The use of Specificity was subject to Θ_{TN} with 98.38% and 96.67% for (a) and (b), respectively. Table 4 compares the evaluation results between Specificity and Sensibility.

was obtained by averaging multiple manual delineations as described in Conformity section. This anatomical structure was segmented by different methods that produced either binary {0, 1} or fuzzy [0, 1] segmentation results as depicted in Figs. 6b–d. To compute the performance measure coefficients, the values of Θ_{FP}, Θ_{FN}, Θ_{TP}, and Θ_{TN} were calculated using Eqs. (A.3) and (A.4) in the Appendix. The global performance evaluation using the Dice, Jaccard, and Conformity coefficients is shown in Fig. 7 and the evaluation based upon the Sensitivity, Specificity, and Sensibility coefficients is presented in Table 2 . It revealed that both Conformity and Sensibility provided relatively wider score ranges and more distinctive comparisons.

To realize more thoroughly the essential characteristics of using the Specificity η_{spf} and Sensibility η_{sbl} coefficients, we compared their reliability in assessing the segmentation of a simulated neuroanatomical structure in synthetic images corrupted with 10% Gaussian noise that were generated in the MATLAB® environment. As shown in Fig. 8, the phantom images with different dimensions 320×320 and 512×512 were segmented using the CFM algorithm (Chang, 2006) and they had the exactly identical results with unanimous scores of $\kappa_c = 39.52\%$ and $\eta_{stv} = 85.45\%$ as depicted in Figs. 8a and b, respectively.

As revealed in Table 3, the scores of Sensibility were consistent with a uniform value of 62.87% for both results. However, the scores of Specificity were different in that the larger the image dimension the higher the score value due to Θ_{TN} [see Eq. (15)].

We further compare the use of Specificity and Sensibility in evaluating the segmentation of neuroanatomical structures in images with different dimensions.

In Fig. 9, we illustrate the segmentation of simulated anatomical structures in images with different dimensions 512×512 and

Table 4
Comparison of using the Specificity and Sensibility coefficients in assessing the performance of the segmentation results in Fig. 9.

Size	Image	Conformity	Sensibility	Specificity
512×512	Fig. 9a	−182.19%	−156.17%	98.38%
256×256	Fig. 9b	60.13%	78.91%	96.67%

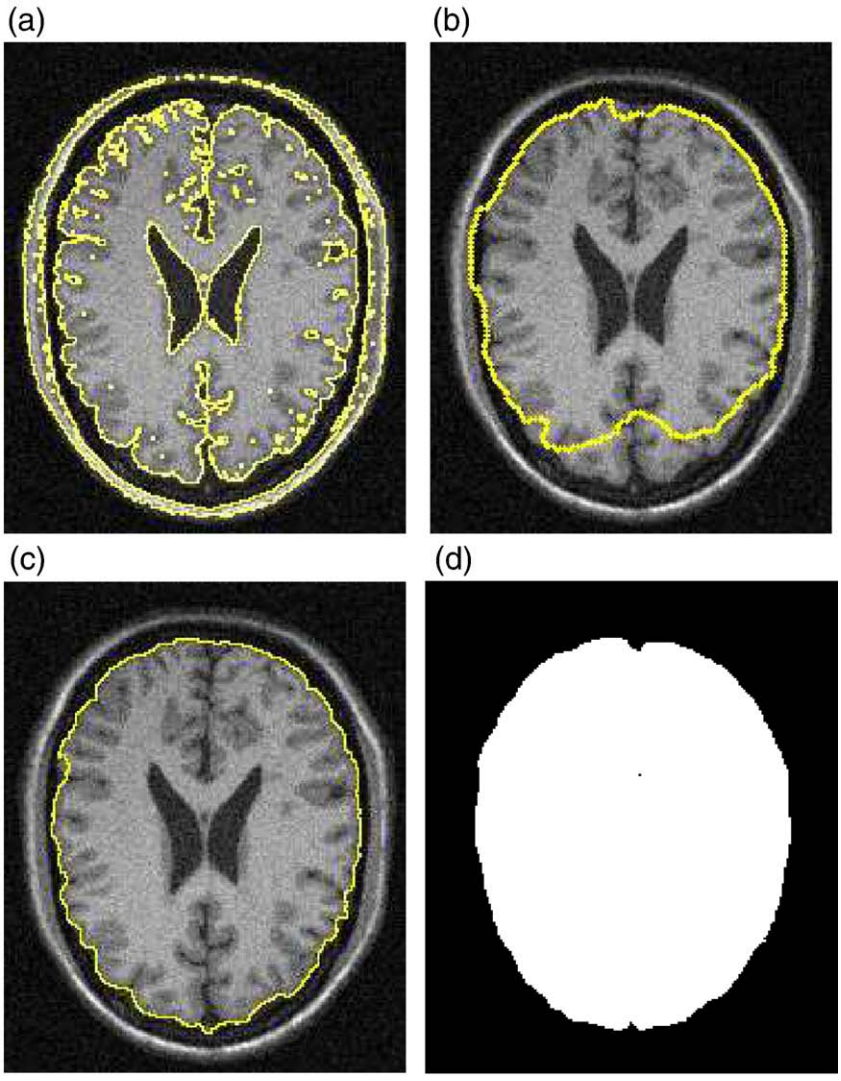

Fig. 10. Segmentation of simulated T1-weighted brain MR images (181×217) of slice 95 with 9% noise and 40% intensity non-uniformity obtained from the BrainWeb simulator. (a) Results of using the CVL method. (b) Results of using the ACM algorithm. (c) Results of using the CFM algorithm. (d) The reference mask obtained from the BrainWeb simulator. Fig. 11 compares the performance evaluation results between the Dice, Jaccard, and Conformity coefficients.

256×256 using the ACM method (Kass et al., 1988). It is observed that the result in Fig. 9a was worse than Fig. 9b in terms of over-segmentation. Nevertheless, the score value of Specificity in Fig. 9a was higher than Fig. 9b, 98.38% vs. 96.67%, as presented in Table 4. This is because a small structure in a larger image has a higher Θ_{TN}. It was this higher Θ_{TN} that contributed to a higher score of Specificity for Fig. 9a. On the other hand, as consistent with the observation the scores of Sensibility were −156.17% and 78.91% for Figs. 9a and b, respectively.

To illustrate the properties of the Dice, Jaccard, and Conformity coefficients, we used the CVL, CML, ACM, and CFM algorithms to segment a variety of brain MR images. In Fig. 10, we show the brain extraction results on the T1-weighted brain MR image obtained from slice 95 of the BrainWeb simulator repository (Kwan et al., 1999) with 9% noise and 40% intensity non-uniformity that had a dimension equal to 181×217. Note that, as depicted in Fig. 10a, the CVL method separated the image into two distinct regions with different intensity distributions since it was designed to be a 2-phase segmentation algorithm. Comparison of the global performance measure scores for Fig. 10 using the Dice, Jaccard, and Conformity coefficients is shown in Fig. 11, which indicates that the Conformity coefficient had approximately 2.1 and 3.1 times wider score ranges than the Jaccard and Dice coefficients, respectively.

Fig. 12 shows the brain segmentation results of female AD images obtained from the CCB Test Data Archive in the LONI at UCLA with a dimension equal to 181×217 using the ACM, CML, and CFM methods. For a more intensive evaluation, we compared Conformity with six other coefficients in Table 1. As illustrated in Fig. 13, the score range of Conformity was approximately 1.9, 2.9, 3.3, and 3.8 times wider than the score ranges of the Jaccard, Dice, Ochiai, and Kulczynski

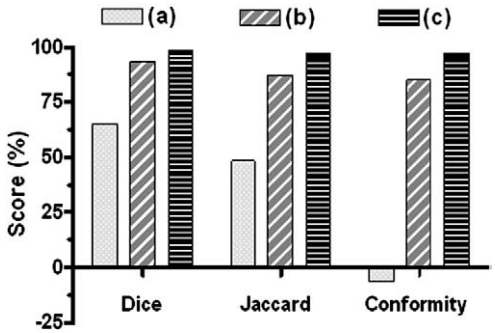

Fig. 11. Comparison of the performance evaluation for Fig. 10 between the Dice, Jaccard, and Conformity coefficients, whose score ranges were 33.5%, 49.1%, and 103.9%, respectively.

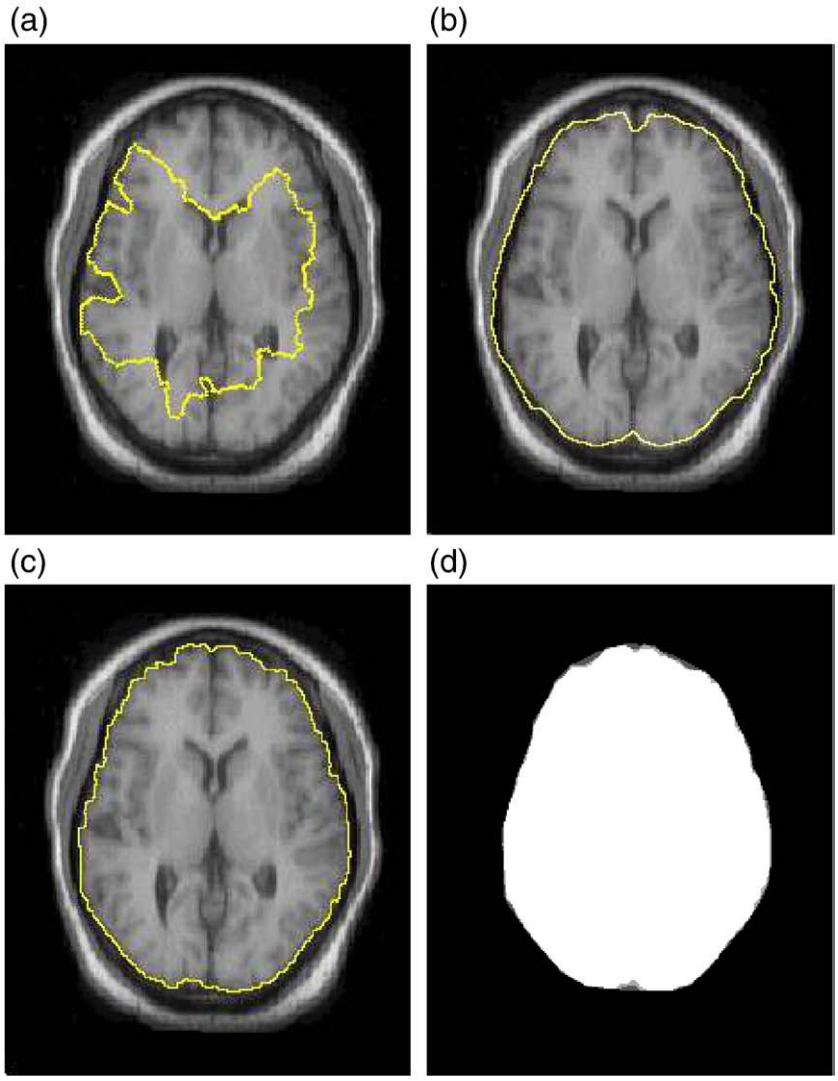

Fig. 12. Segmentation of brain MR images (181 × 217) with AD. (a) Results of using the ACM algorithm. (b) Results of using the CML method. (c) Results of using the CFM algorithm. (d) The fuzzy reference. The performance evaluation of using the Dice, Jaccard, and Conformity coefficients is shown in Fig. 13.

coefficients, respectively. In Fig. 14, we show the segmentation results of brain tumors with blurred boundaries in 256 × 256 T2-weighted MR images obtained from the medical image database in the DINR at UCLA using the CVL, CML, and CFM algorithms. As depicted in Fig. 15, Jaccard, Dice, and Blanque coefficients had quite similar performance measure values and score ranges. The Conformity coefficient provided much better discrimination abilities with a score range of 403.1% that was approximately 8.8, 11.2, and 20.5 times wider than Jaccard, Ochiai, and Kulczynski, respectively. Note that, in both illustrations, the score values of Simpson were somewhat incorrect and the score ranges were extremely narrow as compared to other coefficients. The reference masks in both Figs. 12 and 14 were with fuzzy values, which were obtained by averaging three to four manual delineations by experts.

We conclude this section by illustrating the merits of Conformity and Sensibility to evaluate different automated segmentation techniques in skull stripping 20 normal brain MR image volumes obtained from the IBSR (MGH, 2007). Each of the image volumes had approximately 60 slices with 256 × 256 pixels per slice. The experiments were executed using the BET (Smith, 2002), BSE (Shattuck et al., 2001), and MLS (Zhuang et al., 2006) methods. Fig. 16 indicates that the Conformity coefficient was more sensitive to variations in segmented data with much wider score ranges and larger standard deviations. The range of mean score values for Conformity was approximately 2.0, 2.5, and 3.5 times wider than the Anderberg, Jaccard, and Dice coefficients, respectively. Similarly, as illustrated in Fig. 17, the Sensibility coefficient provided better discrimination

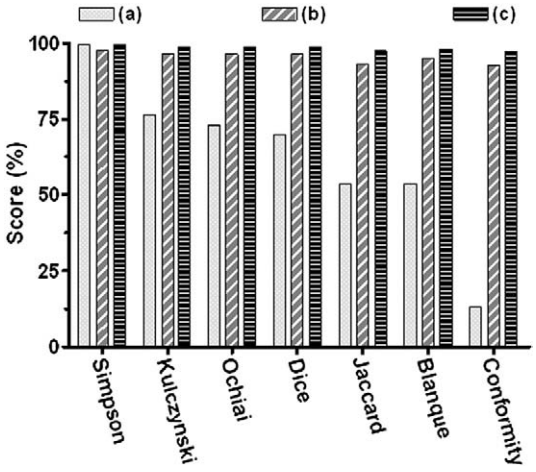

Fig. 13. Comparison of the performance evaluation for Fig. 12 between the Simpson, Kulczynski, Ochiai, Dice, Jaccard, Blanque, and Conformity coefficients, whose score ranges were 1.7%, 22.1%, 25.6%, 29.0%, 43.9%, 44.4%, and 84.2%, respectively.

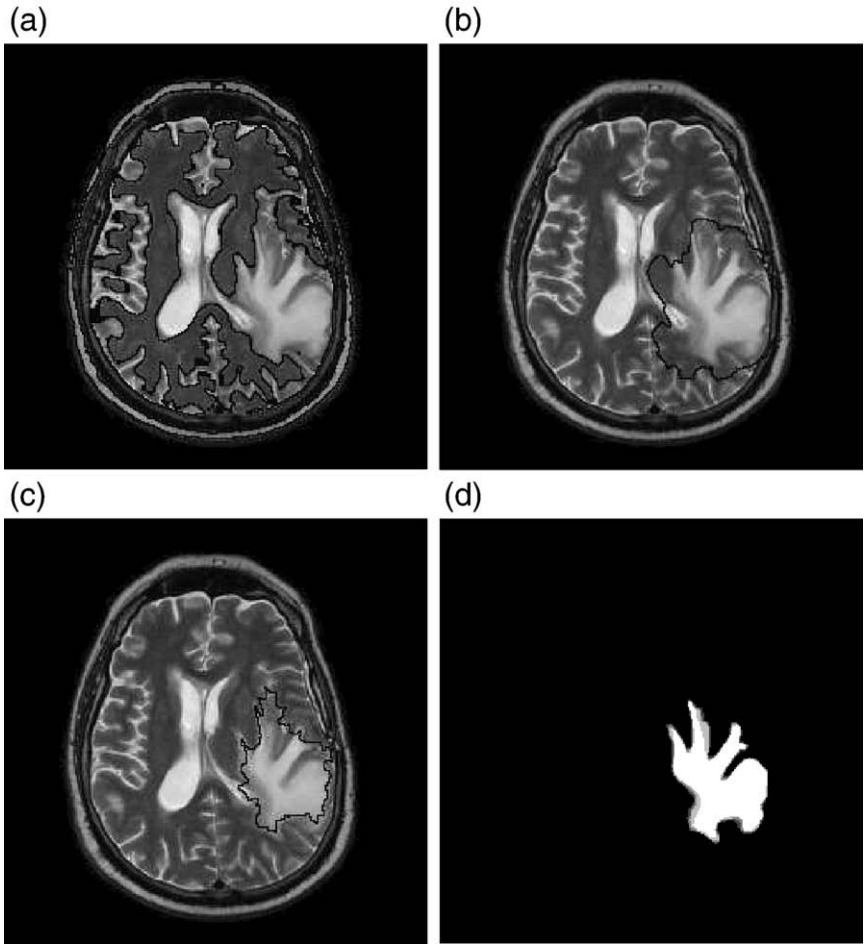

Fig. 14. Segmentation of brain tumors with blurred boundaries in 256×256 T2-weighted MR images. (a) Results of using the CVL method. (b) Results of using the CML algorithm. (c) Results of using the CFM algorithm. (d) The fuzzy reference. Fig. 15 compares the performance evaluation results between the Dice, Jaccard, and Conformity coefficients.

abilities with an approximately 10.9 times wider mean score range than Specificity. Note that Conformity and Sensibility both coefficients not only produced more distinctive evaluation scores but they also enhanced the diversity that resulted in larger standard deviations for each method.

Discussion and conclusions

The objective of this study was to investigate the intrinsic properties and essential characteristics of the fundamental performance measure coefficients that have been extensively used as fundamental elements in many evaluation frameworks for assessing neuroimage segmentation algorithms. In the meantime, we developed a global performance measure coefficient Conformity based upon the spatial overlap of the fuzzy segmentation Ω_{fs} and the fuzzy reference Ω_{fr}. This new global performance measure coefficient can be interpreted from the perspective of measuring the ratio of the mis-segmented region Θ_{AE} to the correctly segmented region Θ_{TP}, which was defined as the discrepancy-to-concordance ratio ξ as shown in Eqs. (3) and (6). While the relationship of Dice to Jaccard has been described in the literature (Shattuck et al., 2001; Crum et al., 2006), e.g., Eqs. (9a) and (9b), their properties have rarely been exploited explicitly and their absolute values are difficult to interpret (Zijdenbos

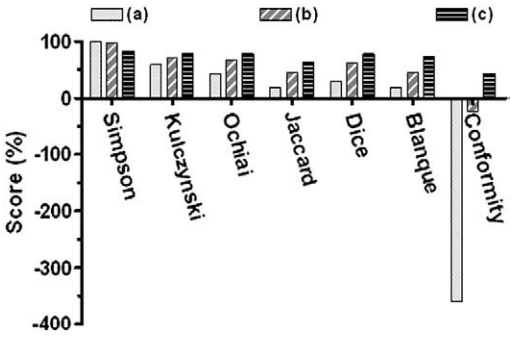

Fig. 15. Comparison of the performance evaluation for Fig. 14 between the Simpson, Kulczynski, Ochiai, Jaccard, Dice, Blanque, and Conformity coefficients, whose score ranges were 16.2%, 19.7%, 36.0%, 46.0%, 47.6%, 55.5%, and 403.1%, respectively.

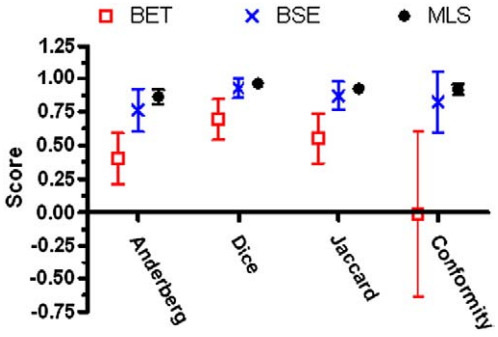

Fig. 16. Comparison between the Anderberg, Dice, Jaccard, and Conformity coefficients to evaluate the BET, BSE, and MLS methods in skull stripping 20 normal brain MR image volumes obtained from the IBSR. The ranges of the mean scores were 26.75%, 37.42%, 45.88%, and 92.97% for Dice, Jaccard, Anderberg, and Conformity, respectively.

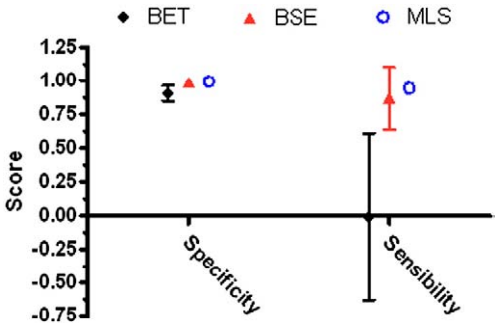

Fig. 17. Comparison between the Specificity and Sensibility coefficients in evaluating the skull stripping results in Fig. 16. The ranges of the mean scores were 8.75% and 95.37% for Specificity and Sensibility, respectively.

et al., 1994). As illustrated in Fig. 2, we explored the score distributions of the Dice and Jaccard coefficients with respect to Θ_{AE} and Θ_{TP} and established the mathematical relationship among Dice, Jaccard, and Conformity.

The merits of Conformity were illustrated by evaluating four different computerized methods in segmenting brain MR images as depicted in Figs. 10–15. Mathematically, the score range of Conformity is $(-\infty, 1]$ based upon Eq. (3), while other similarity coefficients are bounded between 0 and 1, see Table 1. As consistent with Fig. 4, we observed that the score values of κ_c were always smaller than the scores of κ_j and κ_d in the illustrations. The Conformity coefficient provided a much wider score range than the Jaccard and Dice coefficients, especially when the segmentation results were worse. This can be interpreted from Fig. 3, which indicates that Conformity has better discrimination capabilities based upon ξ. Accordingly, Conformity is better able to detect small variations in segmented images.

One particular characteristic of the Specificity coefficient is the incorporation of Θ_{TN} into the computation as shown in Eq. (15). Consequently, Specificity is used more appropriately in detection and classification applications when both foreground and background structures are important to the goal of segmentation. In situations of extracting specific brain structures, e.g., the corpus callosum, hippocampus, and cortex, the background associated with Θ_{TN} is usually meaningless and unimportant for further analyses. The use of Specificity, which is subject to Θ_{TN}, to evaluate such segmentation results could be insensitive and inaccurate.

First, identical results in images with different dimensions have consistent Conformity, Jaccard, Dice, and Sensitivity values but different Specificity scores as illustrated in Fig. 8 and Table 3. Statistical analysis of Specificity for a large number of segmentation results in images with different dimensions may introduce extra variances due to Θ_{TN}. In addition, Conformity, Jaccard, or Dice is usually associated with Sensitivity and Specificity for a more thorough evaluation. Different values of Specificity are in conflict with other performance measure coefficients, whose score values are unanimous. Second, it is possible that the score of Specificity for poor results is higher than the score of better results, e.g., objects with different dimensions in various images as illustrated in Fig. 9 and Table 4. Third, the η_{spf} scores tend to be high and close to one another due to Θ_{TN}, which is usually large, i.e., $\eta_{spf} \approx 1$ when $\Theta_{TN} \gg \Theta_{FP}$. Consequently, it is not sensitive to evaluate small neuroanatomical structures in large images in that statistical analysis of Specificity may not provide apparent differences.

We addressed this by proposing a new measure coefficient Sensibility, whose score values do not rely on Θ_{TN}. Sensibility can be interpreted by measuring the ratio of the over-segmented region Θ_{FP} to the fuzzy reference region Ω_{fr}. The maximum score of η_{sbl} is 100% when $\Theta_{FP} = 0$, a zero score of η_{sbl} is obtained when $\Theta_{FP} = \Omega_{fr}$, and negative scores are obtained when $\Theta_{FP} > \Omega_{fr}$. Comparing to Specificity, Sensibility provides more distinctive and reliable scores in assessing the performance of segmentation algorithms based upon Θ_{FP} as illustrated in Fig. 17 and Table 4. In addition, it is interesting to note that $\eta_{sbl} = \eta_{stv} = \frac{\Theta_{TP}}{\Omega_{fr}}$ when $\Theta_{FP} = \Theta_{FN}$ [see Eqs. (14) and (16)]. In other words, the scores of Sensitivity and Sensibility are all based upon Ω_{fr} and they are identical when Θ_{FN} and Θ_{FP} are equal.

Although the Jaccard, Dice, and Specificity coefficients provided differences to characterize the performance of segmentation algorithms, it was possible that the evaluation scores were statistically alike with a narrow score range as illustrated in Results section. Consequently, it was not easy to rigorously distinguish the performance of different segmentation algorithms, especially when the results were poor. For the purpose of statistical inferences, Zou et al. (2004) proposed a logit transformation of the Dice coefficient using $\text{logit}(\kappa_d) = \ln\{\kappa_d/(1-\kappa_d)\}$. This logarithmic transformation maps the domain of κ_d from $[0, 1]$ to the unbounded range $(-\infty, \infty)$ for better evaluation. On the other hand, the domain of the proposed coefficients Conformity and Sensibility is inherently unbounded in the range $(-\infty, 1]$ that provides more discriminative score values for interpretation.

In summary, a new global performance measure coefficient Conformity was proposed and its merits were described and

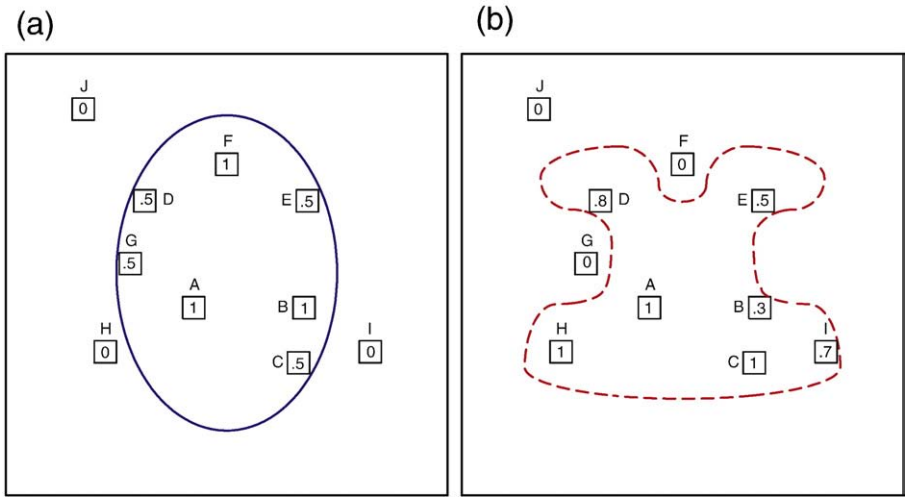

Fig. A1. Schematic illustrating the computation of fuzzy regions for evaluation. (a) Λ_{fr}: the mask of the fuzzy reference. (b) Λ_{fs}: the mask of the fuzzy segmentation.

illustrated. We established the mathematical relationship and explored the intrinsic properties among the Conformity, Jaccard, and Dice coefficients. Conformity had better discrimination capabilities in detecting small variations in segmented images. A new evaluation coefficient Sensibility was also proposed to measure the error of mis-segmented pixels (or voxels) outside the gold standard. Comparing to Specificity, Sensibility had consistent and reliable scores in evaluating the over-segmentation error, especially for small neuroanatomical structures. This study provided some evidence for the merits of Conformity and Sensibility, which can be further incorporated into more sophisticated evaluation frameworks with statistical and variational analyses in a wide variety of neuroimage processing applications.

Acknowledgments

This work was supported in part by the NIH/NIMH research grant No. U54 RR021813 entitled Center for Computational Biology (CCB) and the National Science Council under Research Grant No. NSC96-2811-E-010-001.

Appendix A. Computation of fuzzy regions

To accommodate situations in that both reference and segmentation masks are fuzzy regions, we describe the computation of Θ_{TP}, Θ_{TN}, Θ_{FP}, and Θ_{FN} based upon the fuzzy labeling, which takes values in the continuous domain [0, 1]. Let Λ_{fr} be the mask of the fuzzy reference and Λ_{fs} be the mask of the fuzzy segmentation as illustrated in Fig. A1. The values in Λ_{fr} and Λ_{fs} are normalized to the range between 0 and 1 using the methods described in Conformity section. We denote the fuzzy values as λ_{fr} and λ_{fs} for an arbitrary position i in Λ_{fr} and Λ_{fs}, respectively. The goal is to compute the agreement between these two masks through a point-by-point comparison. Let us start by discussing some simple cases, where λ_{fr} and λ_{fs} take extreme values, i.e., 0 or 1, corresponding to the binary case. In this scenario, any position with $\lambda_{fr}=\lambda_{fs}=1$ contributes to Θ_{TP}, e.g., A, and with $\lambda_{fr}=\lambda_{fs}=0$ to Θ_{TN}, e.g., J. The realm of Θ_{FP} consists of points with $\lambda_{fr}=0$ and $\lambda_{fs}=1$, e.g., H, and Θ_{FN} with $\lambda_{fr}=1$ and $\lambda_{fs}=0$, e.g., F.

With this philosophy in mind, we now consider situations with either λ_{fr} or λ_{fs} being floating numbers between 0 and 1. For positions with equal values of λ_{fr} and λ_{fs}, e.g., E, they contribute to Θ_{TP} with full scores, since there is no difference between them. However, for positions with different values of $\lambda_{fr}>0$ and $\lambda_{fs}>0$, there should be a fraction contributing to Θ_{TP} and the remaining to either Θ_{FP} or Θ_{FN}, depending on the individual value. To better understand the computation for complicated scenarios of fuzzy evaluation, we define the disparity δ between the pair λ_{fr} and λ_{fs} at an arbitrary position i as

$$\delta_i = |\lambda_{fs}-\lambda_{fr}|_i \tag{A.1}$$

where $|\cdot|$ represents the absolute value. We further define Γ_1 be the set whose fuzzy values in Λ_{fs} are larger than the corresponding values in Λ_{fr} and Γ_2 vice versa:

$$\begin{aligned}\Gamma_1 &= \{i\in M\times N \mid \lambda_{fs}>\lambda_{fr}\}\\ \Gamma_2 &= \{i\in M\times N \mid \lambda_{fr}>\lambda_{fs}\}\end{aligned} \tag{A.2}$$

where M and N represent the width and height of an image under consideration, respectively. For example, $D\in\Gamma_1$ and $B\in\Gamma_2$. Based upon previous discussion, δ contributes to false estimations in terms of either Θ_{FP} or Θ_{FN} that depends on the relative values of λ_{fr} and λ_{fs}. Accordingly, the computation of Θ_{FP} and Θ_{FN} is derived as given in the following equations.

$$\begin{aligned}\Theta_{FP} &= \textstyle\sum_{i\in\Gamma_1}\delta_i = \sum_{i\in\Gamma_1}(\lambda_{fs}-\lambda_{fr})_i\\ \Theta_{FN} &= \textstyle\sum_{i\in\Gamma_2}\delta_i = \sum_{i\in\Gamma_2}(\lambda_{fr}-\lambda_{fs})_i\end{aligned}. \tag{A.3}$$

Note that when $\lambda_{fr}=\lambda_{fs}$ there is no contribution to either Θ_{FP} or Θ_{FN}. On the other hand, the value of $1-\delta$ contributes to either Θ_{TP} or Θ_{TN} that depends on the absolute values of λ_{fr} and λ_{fs} as

$$\begin{aligned}\Theta_{TP} &= \textstyle\sum_{i\in\Gamma_3}(1-\delta)_i\\ \Theta_{TN} &= \textstyle\sum_{i\in\Gamma_4}(1-\delta)_i\end{aligned} \tag{A.4}$$

where Γ_3 and Γ_4 are defined as

$$\begin{aligned}\Gamma_3 &= \{i\in M\times N \mid \lambda_{fs}>0 \quad \text{and} \quad \lambda_{fr}>0\}\\ \Gamma_4 &= \Gamma_3^c\end{aligned} \tag{A.5}$$

where Γ_4 is the complement of Γ_3. Consequently, Eqs. A.3 and A.4 can be used for computing the performance measure coefficients described in Methodology section in both fuzzy and binary scenarios.

References

Ali, A.A., Dale, A.M., Badea, A., Johnson, G.A., 2005. Automated segmentation of neuroanatomical structures in multispectral MR microscopy of the mouse brain. NeuroImage 27 (2), 425–435.
Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26 (3), 839–851.
Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1 (8476), 307–310.
Caselles, V., Catte, F., Coll, T., Dibos, F., 1993. A geometric model for active contours in image processing. Numer. Math. 66, 1–31.
Chalana, V., Kim, Y., 1997. A methodology for evaluation of boundary detection algorithms on medical images. IEEE Trans. Med. Imag. 16 (5), 642–652.
Chan, T.F., Vese, L.A., 2001. Active contours without edges. IEEE Trans. Image Process. 10 (2), 266–277.
Chang, H.-H., 2006. Medical image segmentation using an electrostatic charged fluid model. Ph.D. thesis, UCLA.
Cox, T.F., Cox, M.A., 2000. Multidimensional Scaling, 2nd Edition. Chapman & Hall/CRC, Boca Raton.
Crum, W.R., Camara, O., Hill, D.L., 2006. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans. Med. Imag. 25 (11), 1451–1461.
Dawant, B.M., Hartmann, S.L., Thirion, J.-P., Maes, F., Vandermeulen, D., Demaerel, P., 1999. Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations: Part I, methodology and validation on normal subjects. IEEE Trans. Med. Imag. 18 (10), 909–916.
Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.
Dogdas, B., Shattuck, D.W., Leahy, R.M., 2005. Segmentation of skull and scalp in 3-D human MRI using mathematical morphology. Hum. Brain Mapp. 26 (4), 273–285.
Donner, A., Zou, G., 2002. Interval estimation for a difference between intraclass kappa statistics. Biometrics 58 (1), 209–215.
Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.
Fenster, A., Chiu, B., 2005. Evaluation of segmentation algorithms for medical imaging. Conf. Proc. IEEE Eng. Med. Biol. Soc. 7, 7186–7189.
Fiez, J.A., Damasio, H., Grabowski, T.J., 2000. Lesion segmentation and manual warping to a reference brain: intra- and interobserver reliability. Hum. Brain Mapp. 9 (4), 192–211.
Frangi, A.F., Niessen, W.J., Hoogeveen, R.M., van Walsum, T., Viergever, M.A., 1999. Model-based quantitation of 3-D magnetic resonance angiographic images. IEEE Trans. Med. Imag. 18 (10), 946–956.
Haralick, R.M., 1994. Performance characterization in computer vision. CVGIP: Image Understanding 60 (2), 245–249.
Heinonen, T., Dastidar, P., Frey, H., Eskola, H., 1999. Applications of MR image segmentation. Int. J. Bioelectromagn. 01 (01), 35–46.
Hernandez, M., Frangi, A.F., 2007. Non-parametric geodesic active regions: method and evaluation for cerebral aneurysms segmentation in 3DRA and CTA. Med. Image Anal. 11 (3), 224–241.
Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15 (9), 850–863.
Jaccard, P., 1912. The distribution of flora in the alpine zone. New Phytol. 11 (2), 37–50.
Joshi, A.A., Shattuck, D.W., Thompson, P.M., Leahy, R.M., 2007. Surface-constrained volumetric brain registration using harmonic mappings. IEEE Trans. Med. Imag. 26 (12), 1657–1669.
Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: active contour models. Int. J. Comput. Vis. 01 (04), 321–331.
Kaus, M.R., Warfield, S.K., Nabavi, A., Black, P.M., Jolesz, F.A., Kikinis, R., 2001. Automated segmentation of MR images of brain tumors. Radiology 218 (2), 586–591.
Kloppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J., 2008. Automatic classification of MR scans in Alzheimers disease. Brain 131 (3), 681–689.
Kwan, R.K.-S., Evans, A.C., Pike, G.B., 1999. MRI simulation-based evaluation of image-processing and classification methods. IEEE Trans. Med. Imag. 18 (11), 1085–1097.
MacDonald, D., Kabani, N., Avis, D., Evans, A.C., 2000. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. NeuroImage 12 (3), 340–356.
Malladi, R., Sethian, J.A., Vemuri, B.C., 1995. Shape modeling with front propagation: a level set approach. IEEE Trans. Pattern Anal. Mach. Intell. 17 (02), 158–175.

MGH, 2007. Internet Brain Segmentation Repository (IBSR). http://www.cma.mgh.harvard.edu/ibsr/.
Pham, D.L., Xu, C., Prince, J.L., 2000. Current methods in medical image segmentation. Annu. Rev. Biomed. Eng. 02, 315–337.
Powell, S., Magnotta, V.A., Johnson, H., Jammalamadaka, V.K., Pierson, R., Andreasenb, N.C., 2008. Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. NeuroImage 39 (1), 238–247.
Shan, Z.Y., Yue, G.H., Liu, J.Z., 2002. Automated histogram-based brain segmentation in T1-weighted three-dimensional magnetic resonance head images. NeuroImage 17 (3), 1587–1598.
Sharief, A.A., Badea, A., Dale, A.M., Johnson, G.A., 2008. Automated segmentation of the actively stained mouse brain using multi-spectral MR microscopy. NeuroImage 39 (1), 136–145.
Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13 (5), 856–876.
Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.
Suri, J.S., Setarehdan, S.K., Singh, S., 2002. Advanced algorithmic approaches to medical image segmentation: state-of-the-art applications in cardiology, neurology, mammography and pathology. Springer-Verlag, London.
Toga, A.W., Mazziotta, J.C., 2002. Brain Mapping the Methods, 2nd Edition. Academic Press, San Diego.
Udupa, J.K., LeBlanc, V.R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L.M., Hirsch, B.E., Woodburn, J., 2006. A framework for evaluating image segmentation algorithms. Comput. Med. Imag. Grap. 30 (2), 75–87.
Vannier, M.W., Pilgram, T.K., Speidel, C.M., Neumann, L.R., Rickman, D.L., Schertz, L.D., 1991. Validation of magnetic resonance imaging (MRI) multispectral tissue classification. Comput. Med. Imag. Grap. 15 (4), 217–223.
Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23 (7), 903–921.
Wu, Y., Warfield, S.K., Tan, I.L., Wells, W.M., Meier, D.S., van Schijndel, R.A., Barkhof, F., Guttmann, C.R., 2006. Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI. NeuroImage 32 (3), 1205–1215.
Zhang, Y.J., 1996. A survey on evaluation methods for image segmentation. Pattern Recogn. 29 (8), 1335–1346.
Zhuang, A.H., Valentino, D.J., Toga, A.W., 2006. Skull-stripping magnetic resonance brain images using a model-based level set. NeuroImage 32 (1), 79–92.
Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C., 1994. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans. Med. Imag. 13 (4), 716–724.
Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M., Kaus, M.R., Haker, S.J., Wells, W.M., Jolesz, F.A., Kikinis, R., 2004. Statistical validation of image segmentation quality based on a spatial overlap index. Acad. Radiol. 11 (2), 178–189.