ELSEVIER

Computerized Medical Imaging and Graphics 30 (2006) 75–87

Computerized Medical Imaging and Graphics

www.elsevier.com/locate/compmedimag

A framework for evaluating image segmentation algorithms

Jayaram K. Udupa [a,*], Vicki R. LeBlanc [b,c], Ying Zhuge [a], Celina Imielinska [b,d,e], Hilary Schmidt [b,c], Leanne M. Currie [f], Bruce E. Hirsch [g], James Woodburn [h]

[a] *Medical Image Processing Group, Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA*
[b] *Office of Scholarly Resources, Columbia University College of Physicians and Surgeons, New York, NY, USA*
[c] *Center for Education Research and Evaluation, Columbia University College of Physicians and Surgeons, New York, NY, USA*
[d] *Department of Biomedical Informatics, Columbia University College of Physicians and Surgeons, New York, NY, USA*
[e] *Department of Computer Science, Columbia University, New York, NY, USA*
[f] *School of Nursing, Columbia University, New York, NY, USA*
[g] *Department of Neurobiology and Anatomy, Drexel University College of Medicine, Philadelphia, PA, USA*
[h] *Rheumatology and Rehabilitation Research Unit, University of Leeds, Leeds, UK*

Received 27 August 2005; accepted 12 December 2005

Abstract

The purpose of this paper is to describe a framework for evaluating image segmentation algorithms. Image segmentation consists of object recognition and delineation. For evaluating segmentation methods, three factors—precision (reliability), accuracy (validity), and efficiency (viability)—need to be considered for both recognition and delineation. To assess precision, we need to choose a figure of merit, repeat segmentation considering all sources of variation, and determine variations in figure of merit via statistical analysis. It is impossible usually to establish true segmentation. Hence, to assess accuracy, we need to choose a surrogate of true segmentation and proceed as for precision. In determining accuracy, it may be important to consider different 'landmark' areas of the structure to be segmented depending on the application. To assess efficiency, both the computational and the user time required for algorithm training and for algorithm execution should be measured and analyzed. Precision, accuracy, and efficiency factors have an influence on one another. It is difficult to improve one factor without affecting others. Segmentation methods must be compared based on all three factors, as illustrated in an example wherein two methods are compared in a particular application domain. The weight given to each factor depends on application.

Keywords: Image segmentation; Evaluation of segmentation; Image analysis; Segmentation efficacy

1. Introduction

1.1. Background

Image segmentation is the process of identifying and delineating objects in images. It is the most crucial among all computerized operations done on acquired images. Even seemingly unrelated operations like image (gray-scale/color) display, 3D visualization, interpolation, filtering, and registration depend to some extent on image segmentation since they all would need some object information for their optimum performance. Ironically, segmentation is needed for segmentation itself since object knowledge facilitates segmentation. In spite of several decades of research [1,2], segmentation remains a challenging problem in image processing and computer vision.

Image segmentation may be thought of as consisting of two related processes—*recognition* and *delineation*. *Recognition* is the high-level process of determining roughly the whereabouts of an object of interest in the image. *Delineation* is the low-level process of determining the precise spatial extent and point-by-point composition (material membership percentage) of the object in the image. Humans are more qualitative and less quantitative, whereas, computerized algorithms are more quantitative and less qualitative. Incorporation of high-level expert

* Corresponding author. Address: Medical Image Processing Group, Department of Radiology, University of Pennsylvania, 423 Guardian Drive, Fourth Floor, Blockley Hall, Philadelphia, PA 19104-6021, USA. Tel.: +1 215 662 6780; fax: +1 215 898 9145.
E-mail address: jay@mipg.upenn.edu (J.K. Udupa).

0895-6111/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compmedimag.2005.12.001

human knowledge algorithmically into the computer has remained a challenge. Most of the drawbacks of current segmentation methods may thus be attributed to the latter weakness of computers in the recognition process. We envisage, therefore, that the assistance of humans, knowledgeable in the application domain, will remain essential in any practical image segmentation method. The challenge and goal for image scientists are to develop methods that minimize the degree of this required help as much as possible.

While algorithms for image segmentation have been in development for several decades [1,2], the development of systematic evaluation frameworks for these algorithms has been lagging, particularly in medical imaging which is the focus of this paper. The lag is perhaps the result of problems such as limits in common data sets with which to compare methods, difficulty in defining the performance metrics and statistics, and the difficulty in establishing true segmentation. As early as 1977, the need for effective evaluation of the segmentation of biological images has been outlined [3]. More recently, this need has been echoed by many researchers [2,4–6]. In [4,5], the authors stress the need for an objective evaluation of medical image segmentation on large sets of common clinical data, arguing that this is a critical step towards establishing the validity and the clinical applicability of an algorithm. Similarly, [6] claims that the development of an objective approach will provide consistency in evaluation methods by removing biases due to human factors. Many attempts at evaluation do not address the important components that should be present in any evaluation methodology, thus limiting their validity and clinical applicability. Claims about the performance of segmentation algorithms are limited by problems such as (a) the data sets are too small, (b) different data sets are used for different estimations of performance, (c) the data sets are not representative of a clinical problem, (d) appropriate ground truths (or surrogates) are difficult to determine, (e) the performance metrics are poorly defined, (f) there is poor methodology for training and testing the algorithms, (g) large costs of time and effort are involved in collecting and hand-segmenting data, and (h) the algorithms are not compared against other algorithms [5,7].

In light of such difficulties, it is not surprising that many researchers develop complex applications (e.g. virtual colonoscopy systems) that make use of 3D visualizations of anatomical images derived from 3D segmentation methods that have not been formally evaluated by a consistent evaluation strategy (e.g. [8,9]). Many of the researchers who do evaluate their segmentation algorithms do so only on a limited number of components, such as cost analysis [10], inter-rater reliability [11], overall volume [12], or the Hausdorff distance [13]. These efforts, despite representing a valid attempt at evaluation, exemplify the difficulty in devising comprehensive and effective segmentation evaluation methodologies in this domain.

Few researchers [4–7] have made attempts to develop evaluation frameworks that incorporate many of the performance metrics necessary for a practical and informative evaluation of a segmentation algorithm. In [7], the authors discuss the variety of metrics that would result in a valid estimation of the performance of an algorithm. When comparing a segmentation method to a ground truth segmentation of the image, [7] argues that there are five possible outcomes that need to be identified. The computer algorithm can either (a) correctly segment a region, (b) over-segment a region, (c) under-segment a region, (d) miss a region, or (e) incorrectly segment a noise region. Hoover et al. [5] also developed a rigorous framework for the evaluation of segmentation algorithms. This involved the use of pixel-level ground truths in 30 real images. The ground truth consisted of the hand-segmentation, which was reviewed by a second human operator to catch obvious errors. Each pixel in the region segmented by the computer algorithm was classified as either a correct detection, an over-segmentation, an under-segmentation, a missed pixel, or noise. Four algorithms were then compared and described on the basis of these metrics, as well as on the basis of processing time. Zhang [6] approaches evaluation of segmentation methods by proposing analytical and empirical methods, where the empirical methods are divided into goodness and discrepancy measurements. The analytical methods examine and assess the segmentation algorithms themselves by analyzing their principles and properties. The empirical methods indirectly judge the algorithms by testing the images and evaluating the quality of segmentation results. The weakness of this approach is that it is intended for 'all images'. Because of the lack of a general theory for image segmentation, not all characteristics of segmentation can be obtained and described by analytical studies.

We argue that a primary reason for the lack of activity in evaluation, commensurate with the level of investigation in segmentation algorithm development, is the lack of a framework which algorithm developers can readily utilize, without having to spend a great deal of time, to assess the efficacy of their methods. Such a framework, we believe, should consist of: (F1) a specification of readily computable, effective, and meaningful metrics of efficacy, (F2) real life image data, (F3) reference segmentations that can be used as surrogates of true segmentations (ground truth), (F4) a few standard segmentation algorithms, and (F5) a software system that incorporates the evaluation methods and the standard segmentation algorithms. We shall use the phrase *evaluation framework* to refer to this quintuple of components (F1)–(F5). It is clear from the above description that a comprehensive framework for the evaluation of segmentation algorithms in the sense of including the five components is lacking. Even the metrics of efficacy have not considered all important

factors and situations that may influence the segmentation results. For example, variations due to images acquired on different imaging devices (brands, sites, etc.) or slight changes in the acquisition protocol are rarely considered. Even the variations arising from repeat acquisitions on the same device are rarely addressed.

There are shortcomings in the metrics of efficacy commonly used. For example, the commonly-used volume of an object of interest is not a good metric since two sets S_1 and S_2 of voxels may constitute very similar volumes, although, as segmentations of the same physical object in two different situations, the sets may differ significantly in terms of the extent of overlap ($S_1 \cap S_2$), and false positive and false negative regions ($S_1 - S_2$ and $S_2 - S_1$). Factors involving efficiency (practical viability of the segmentation method in terms of the extent of the various forms of human and computational help needed) are not at all considered except for one aspect of the computational requirements of the method. As we shall see later on, there are several factors (both human and computational) that influence efficiency that should be considered within the evaluation framework. Factors relating to the quality of segmentation results have not been considered previously. These factors allow us to take into account in the evaluation framework how well certain salient features of the object (e.g. a site of attachment of a ligament, a particular segment of the object boundary), which are considered important for the application for which image segmentation is sought, are captured in the segmentation.

1.2. Purpose

In summary, the gaps that exist in the currently used evaluation strategies are of two kinds: methodological and resource related. The former represent lapses in the evaluation techniques currently employed. The present paper is an attempt at filling these methodological gaps. The latter pose challenges to the image scientist since the evaluation tasks require considerable resources (multiple data sets with repeat acquisitions and from different sites and brands of imagers and for different applications with known segmentations, software), which most algorithm developers do not possess. We are working toward addressing these issues and nothing further will be said about these issues in this paper. We believe that further work is needed in each of the five components (F1)–(F5) of the framework. The proposed evaluation methodology is described in Section 2 and an example is presented in Section 3 illustrating how the methodology can be utilized in an actual application for comparing methods. Our concluding remarks are stated in Section 4. An early version of this paper was presented at the SPIE Medical Imaging 2002 conference whose proceedings contained that paper [14].

2. The methodology

2.1. Notation and terminology

Any method of evaluation of segmentation algorithms has to, at the outset, specify the *application domain* under consideration. We consider the application domain to be determined by the following three entities.

T: A *task*; example: volume estimation of tumors.
B: A *body region*; example: brain.
P: An *imaging protocol*; example: FLAIR MR imaging with a particular set of parameters.

An evaluation characterizing the efficacy of a particular segmentation method α for a given application domain $\langle T,B,P\rangle$ that signals high performance for α may tell nothing at all about α for a different application domain $\langle T',B',P'\rangle$. For example, a particular algorithm may have high performance in determining the volume of a tumor in the brain on an MR image, but may have a low performance in segmenting a cancerous mass from a mammography scan of a breast. Therefore, evaluation must be performed for each application domain separately. The following additional notations are needed for our description.

Object: A physical object of interest, denoted $\mathcal{O}$, in *B* for which images are acquired; example: brain tumor.
Scene: A 3D (or higher-dimensional) volume image, denoted by $\mathcal{C} = (C,f)$, where *C* is a 3D (or higher-dimensional) rectangular array of *voxels* (short for volume elements), and $f(c)$ denotes the *scene intensity* of any voxel c in *C*. $\mathcal{C}$ may be a *vectorial scene*, meaning that $f(c)$ may be a vector whose components represent several imaged properties. $\mathcal{C}$ is referred to as a *binary scene* if the range of $f(c)$ is $\{0,1\}$.
$S^{\mathcal{X}}$: A set of scenes acquired for the same given application domain $\mathcal{X} = \langle T,B,P\rangle$ for different subjects.

The word 'segmentation' as used in medical imaging has two distinct meanings in two different contexts. The first context is provided by computer aided diagnosis (CAD). *Detection* in this context refers to the act of finding via the given scene an abnormality (such as a lesion) that may exist in *B*. The answer sought in this act is mostly to the query whether or not a particular kind of abnormality (such as a nodule in the lung) that may exist in *B* is portrayed in the scene. If the answer is 'yes', the purpose of the second act of *localization* is to mark on the scene location(s) where the abnormality is determined to be present. The second context for a different meaning for 'segmentation' is provided by a wide-spread and long-standing activity that begs for a name and an acronym of its own. We refer to this for now as CAVA, an acronym for computer aided visualization and

analysis. Briefly, the goal of CAVA is to develop computer methods for aiding humans in *visualizing* the objects in B in their true form, shape, and function, and to quantify the form, shape, and function of these objects. This is usually for the purpose of studying in vivo the normal behavior of an organ system in B or its disease processes in their natural course or the effects of therapy on the disease processes. The nature of the objectives and requirements for segmentation are quite different in CAD and CAVA, as such, we believe that their evaluation strategies must also be different. The problem of evaluation addressed in this paper is as related to segmentation in CAVA.

Segmentation of an object $\mathcal{O}$ in a given scene acquired for an application domain $\langle T,B,P\rangle$ is the process of defining the region/boundary of $\mathcal{O}$ in the given scene. As said previously, it consists of two related tasks—recognition and delineation. Although recognition in CAVA is analogous to detection in CAD, the term *detection* would be inappropriate to describe the role of recognition. For example, if the task is to quantify brain atrophy in studying a neurological disease or its treatment effects, the high-level act of recognizing brain parenchyma and distinguishing it from other objects in the head that are also portrayed in the scene is very different from seeking an answer to the detection task of 'whether or not the brain is there'. Similar comments are applicable to delineation vis-a-vis localization.

We assume that the output of any segmentation algorithm corresponding to a given scene $\mathcal{C}=(C,f)$ is a (hard) set $O\subset C$ of voxels. This set represents the region occupied by (the support of) an object $\mathcal{O}$ of B in C. To accommodate methods that output fuzzy segmentation results, we denote the *fuzzy segmentation* of $\mathcal{O}$ in $\mathcal{C}$ as a scene $\mathcal{C}_O=(C,f_O)$, where, for any $c\in C$, $f_O(c)$ denotes the degree of *objectness* assigned to every voxel c in O by the segmentation method. We shall always denote a hard segmentation in $\mathcal{C}$ of an object $\mathcal{O}$ in B by O and the corresponding fuzzy object by $\mathcal{C}_O$. Our treatment throughout will be general considering fuzzy objects to be the output of segmentation methods. Hard segmentations will become a particular case of this general treatment.

We will use the following operations on fuzzy segmentations. Let $\mathcal{C}_{Ox}=(C,f_{Ox})$, $\mathcal{C}_{Oy}=(C,f_{Oy})$, and $\mathcal{C}_{Oz}=(C,f_{Oz})$ be any fuzzy segmentations defined by the same physical object $\mathcal{O}$ in a scene $\mathcal{C}$. The *cardinality* $|\mathcal{C}_{Ox}|$ of the fuzzy segmentation $\mathcal{C}_{Ox}$ is defined as $|\mathcal{C}_{Ox}|=\sum_{c\in C} f_{Ox}(c)$. Fuzzy set *union* $\mathcal{C}_{Oz}=\mathcal{C}_{Ox}\cup\mathcal{C}_{Oy}$ is defined by, for any $c\in C$, $f_{Oz}(c)=\max(f_{Ox}(c), f_{Oy}(c))$. Fuzzy set *intersection* $\mathcal{C}_{Oz}=\mathcal{C}_{Ox}\cap\mathcal{C}_{Oy}$ is defined by, for any $c\in C$, $f_{Oz}(c)=\min(f_{Ox}(c), f_{Oy}(c))$. Fuzzy set *complement* $\bar{\mathcal{C}}_{Ox}=(C,\bar{f}_{Ox})$ of $\mathcal{C}_{Ox}$ is defined by, for any $c\in C$, $\bar{f}_{Ox}(c)=1-f_{Ox}(c)$.

Fuzzy set *difference* $\mathcal{C}_{Oz}=\mathcal{C}_{Ox}-\mathcal{C}_{Oy}$ is defined by, for any $c\in C$:

$$f_{Oz}(c)=\begin{cases} f_{Ox}(c)-f_{Oy}(c), & \text{if } f_{Ox}(c)-f_{Oy}(c)\geq 0\\ 0, & \text{otherwise.}\end{cases} \quad (1)$$

A fuzzy masking operation $\mathcal{C}_{Oz}=\mathcal{C}_{Ox}\bullet\mathcal{C}_{Oy}$, called *inside*, is defined by, for any $c\in C$:

$$f_{Oz}(c)=\begin{cases} f_{Ox}(c), & \text{if } f_{Oy}(c)\neq 0\\ 0, & \text{otherwise.}\end{cases} \quad (2)$$

Another fuzzy masking operation $\mathcal{C}_{Oz}=\mathcal{C}_{Ox}\circ\mathcal{C}_{Oy}$ called *outside*, is defined by, for any $c\in C$:

$$f_{Oz}(c)=\begin{cases} f_{Ox}(c), & \text{if } f_{Oy}(c)=0\\ 0, & \text{otherwise.}\end{cases} \quad (3)$$

2.2. Outline of methodology

The *efficacy* of any segmentation method M in an application domain $\langle T,B,P\rangle$ is to be measured in terms of three groups of factors: *Precision* (also known as *reliability*), which represents repeatability of segmentation taking into account all subjective actions required in producing the result; *accuracy* (also known as *validity*), which denotes the degree to which the segmentation agrees with truth; *efficiency* (also known as *viability*), which describes the practical viability of the segmentation method. In evaluating segmentation efficacy, both recognition and delineation aspects must be considered. Commonly, only delineation is considered to represent the entire segmentation process. Our methodology attempts to capture both recognition and delineation within the same framework in the factors considered for evaluation.

Our overall approach for the evaluation method consists of the following steps. (1) Establishing true segmentation for delineation. (2) Establishing true segmentation for recognition. (3) Defining metrics for precision. (4) Defining metrics for accuracy. (5) Defining metrics for efficiency. (6) Comparison of segmentation methods by statistical analysis of the metrics generated by the methods on the same set of scenes acquired for a given application domain $\langle T,B,P\rangle$. These steps are described in detail in the following sections.

2.3. Surrogate of truth

For real scenes (patient scenes in medical imaging), since it is impossible to establish absolute true segmentation, some surrogate of truth is needed. In describing this aspect, we will treat the delineation and recognition aspects separately in Sections 2.3.1 and 2.3.2, respectively.

2.3.1. Object delineation

Three possible choices of the surrogate for delineation are outlined below.

(a) Manual delineation. Object boundaries are traced or regions are painted manually by experts (see Fig. 1). Sometimes, it is easier for experts to manually correct the delineation produced by an algorithm. Corresponding to a given set $S^{\mathcal{X}}$ of scenes for the application domain $\mathcal{X}=\langle T,B,P\rangle$, manual delineation in either of these forms

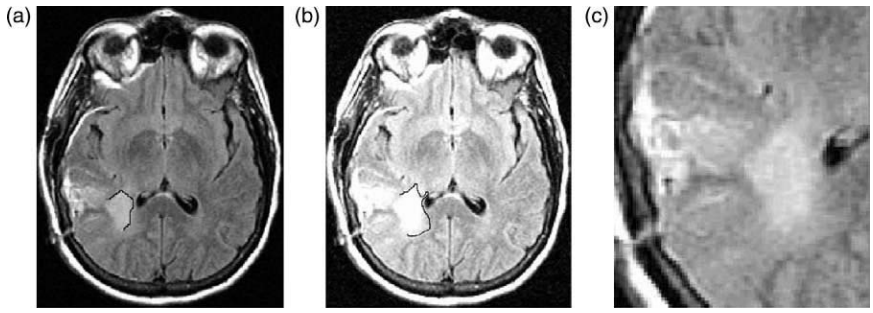

Fig. 1. A slice from an MR FLAIR scene of a patient's brain. Different window settings (a) and (b) and magnification factors (c) can cause significant variations in the result of manual delineations, especially for fuzzy objects.

produces a set $S_{td}^{\mathcal{X}}$ of scenes representing the fuzzy segmentations of the same object represented in the scenes in $S^{\mathcal{X}}$ in the following manner. Manual delineation produces a hard set O for each scene $\mathcal{C}$ in $S^{\mathcal{X}}$. Multiple repetitions of segmentation by multiple operators should be performed. The fuzzy segmentation $\mathcal{C}_O$ is produced simply by averaging the multiple manual delineations. Manual delineation is inherently binary; that is, it cannot specify tissue percentages. These binary results are converted into fuzzy segmentations by using the above strategy. However, if only binary segmentation is desired, then the averaged scene can be thresholded at 0.5 to output a binary scene corresponding to each $\mathcal{C}$ in $S^{\mathcal{X}}$. In that case, $S_{td}^{\mathcal{X}}$ contains binary scenes. Another alternative is to use the method suggested in [15] wherein an expectation–maximization algorithm is described for estimating the surrogate of true delineation produced by a group of experts.

Manual delineation has several shortcomings. First, when required to be done by expert physicians skilled in the application domain, it is very costly because of the time and effort needed in hand segmenting multiple data sets multiple times. Second, it can be highly variable. For example, intra- and inter-operator variations of over 20% have been reported for manual outlining of multiple sclerosis lesions in brain MRI scenes [16,17]. Third, the precision of manual delineations depends on the crispness of boundaries, the window level settings for image display, the computer monitor and its settings, and even on the operator's vision characteristics [18]. When object regions/boundaries are fuzzy or very complex, manual delineation becomes ill-defined. For example, in Fig. 1, it is difficult to decide what aspect of the edematous region of the tumor should be included/excluded. Given the various problems with other surrogates (see below) and given that manual delineation is an accepted standard surrogate, it makes sense to examine how we may overcome some of the drawbacks of manual outlining and still produce a surrogate that is governed by the underlying precept which has made this mode of delineation to be accepted as a defacto standard surrogate. For example, the display characteristics of display monitors can be standardized [19], and the scene intensity scale can be standardized [20] in modalities such as MRI wherein the intensity scales are arbitrary and do not have a tissue-specific numeric meaning. Such strategies make it possible to standardize window settings for scene display to minimize the variation in displayed scenes. These aspects require further work.

(b) Mathematical phantoms. A set of mathematical phantoms is created to depict the application domain $\mathcal{X} = \langle T,B,P\rangle$ as realistically as possible in terms of image blur, relative tissue contrast and heterogeneity, noise, and background inhomogeneity in the scenes (see Fig. 2). The starting point for this simulation is a set $S_{td}^{\mathcal{X}}$ of binary scenes (true delineation is known to begin with). Each scene in $S_{td}^{\mathcal{X}}$ is gradually corrupted to yield the actual set of scenes $S^{\mathcal{X}}$. We may also start with gray scenes depicting true fuzzy segmentations and then follow the same procedure. The approach here is the reverse of that described above for

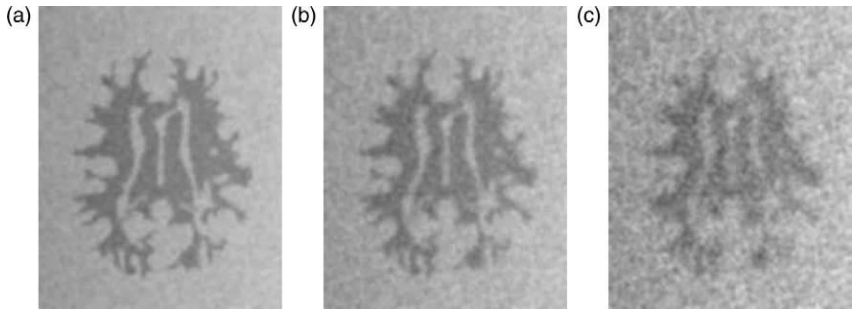

Fig. 2. White matter (WM) in a gray matter background, simulated by segmenting WM from real MR scenes and by setting contrast observed in real scenes and adding blur, noise, background variation to various degree: (a) low, (b) median, and (c) high.

manual delineation in the sense that here we start from $S^{\mathcal{X}}_{td}$ and then produce $S^{\mathcal{X}}$ as compared to producing $S^{\mathcal{X}}_{td}$ starting from $S^{\mathcal{X}}$ in the latter approach. The main shortcoming of this approach is the questionable authenticity of the challenges posed by the scenes in $S^{\mathcal{X}}$ to segmentation algorithms.

(c) Simulated scenes. One of the possible strategies in this approach is to use the method of mathematical phantoms described above to generate scenes and apply to both the segmentations and the simulated scenes known 3D deformations (to capture realistically the variations that exist among patients) to generate more scenes and their segmentations. The same method is applicable to the method of manual segmentation also. The complete set of scenes (original with deformed) in this case constitutes $S^{\mathcal{X}}$, and the complete set of segmentations represents $S^{\mathcal{X}}_{td}$. The main drawback of this approach is that it is difficult to devise deformations and the associated changes in intensity characteristics that are realistic.

Another method is to emulate the process of image acquisition as realistically as possible, starting from mathematical phantoms which constitute object regions, and tissue properties/labels assigned to the regions which constitute actual object properties/labels [21,22]. This process consists of three distinct steps: (a) generating the object geometries; (b) simulating the physical process of data collection to generate the so called 'projection data' based on assumptions regarding the imaging device; (c) reconstructing images. A weakness of this approach is that in step (a), it is very difficult to include all objects in a body region in the mathematical phantom and at sufficiently high resolution. It is also very expensive to generate a sufficient number of data sets corresponding to different imaged subjects. Since, all objects cannot be considered with their realistic properties, realism of the challenges posed for segmentation cannot be guaranteed in the resulting scenes. The object geometries generated in step (a) in this approach constitute $S^{\mathcal{X}}_{td}$, and the reconstructed images generated in step (c) constitute $S^{\mathcal{X}}$.

A third method to simulate scenes is to first create an ensemble of 'cut-outs' of object regions from actual acquired scenes and to bury them realistically in different scenes. Each cutout is segmented carefully by using an appropriate segmentation method. This should not be difficult since the cutout contains just the object region with a background tissue region only and no other confounding tissue regions. The resulting scenes and the segmentations constitute $S^{\mathcal{X}}$ and $S^{\mathcal{X}}_{td}$, respectively. See Fig. 3. A major weakness of this approach is that its applicability is very limited, perhaps to only small objects (such as multiple sclerosis lesions) that occur within large regions of a co-object (such as white matter) in a relatively independent manner.

In our methodology, $S^{\mathcal{X}}_{td}$ and the corresponding set $S^{\mathcal{X}}$ associated with any of the above methods can be utilized. We recommend, however, that the averaged fuzzy results output by the manual delineation method, particularly the method wherein the output of an algorithm is manually corrected by human operators (rather than the results of fully manual delineation) under standardized conditions, be used as a surrogate of true delineation.

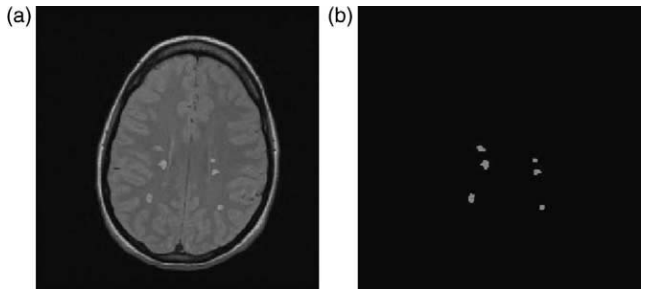

Fig. 3. A slice (a) of a scene simulated from an acquired MRI proton density scene of a multiple sclerosis patient's brain and its 'true' segmentation (b) of the lesions.

2.3.2. Object recognition

Evaluation strategies that are usually considered for assessing delineation accuracy are based on treating all aspects of the region corresponding to the surrogate of true delineation with equal weight. In the absence of any prerequisites, this is a correct, and only possible, stand. However, such approaches do not address the fact that some areas of the object may be more important than others. An algorithm may segment an object and match 98% of the true delineated region. The importance of that 2% difference will depend on the importance of the regions missed in delineation. For example, if it is a crucial landmark area, such as a location of vascularization, or a site of attachment of a ligament on a bone, then missing or overestimating 2% of the volume in this region could have important repercussions for the surgeon or therapist who needs to know the location of vital nearby anatomic objects. This example highlights the importance of landmark identification and weighting in evaluating an algorithm's recognition performance. Our approach for ensuring the inclusion of the information related to certain key features or landmarks related to the object (the recognition aspect) in the surrogate used for assessing accuracy of segmentation is as follows.

(1) Compile a list of features/landmarks (points, curves, regions) on the object that are observable in the acquired scene and that are vital for the application domain $\mathcal{X} = \langle T, B, P \rangle$ through help from a set of experts (radiologists, surgeons, anatomists).
(2) Have each expert assign a score to each feature to indicate its level of importance in the application domain $\mathcal{X}$.
(3) Compute an average of the scores. Normalize these to the range [0, 1]. In this fashion, we generate a feature

vector F whose components have values in [0, 1] indicating their level of importance for $\mathcal{X}$.

(4) Have experts delineate the critical area of these features in scenes in $S^{\mathcal{X}}$ repeatedly.

(5) Use the mean location, the spread information about the location of the features, and the mean vector F to generate a scene $\mathcal{C}^i_{tr} = (C, f^i_{tr})$ for each scene $\mathcal{C} \in S^{\mathcal{X}}$ and for each feature i. $\mathcal{C}^i_{tr}$ is a location-weighted (for feature location) and importance-weighted (for i) representation of feature i. In this scene, a high value of $f^i_{tr}(c)$ for a voxel $c \in C$ indicates that c is both close to the mean location for a particular feature i and the importance of the feature is high. These individual scenes $\mathcal{C}^i_{tr}$ may be combined into a composite scene $\mathcal{C}_{tr}$ by taking an average or a fuzzy union over all i. Fuzzy union is perhaps more appropriate and this is the approach we have taken. From this repeated identification of landmarks/features, we generate the scenes $\mathcal{C}_{tr}$ capturing information about truth in recognition for each scene $\mathcal{C} \in S^{\mathcal{X}}$. Note that $\mathcal{C}_{tr}$ does not have any information about object delineation. It contains bright blobs (of different shapes) only at the location of the selected features. We denote by $S^{\mathcal{X}}_{tr}$ the set of scenes containing 'truth' in recognition for the set of scenes $S^{\mathcal{X}}$.

2.4. Metrics of segmentation efficacy

2.4.1. Assessment of precision

Two types of subjective actions need to be addressed in evaluating segmentation precision: (1) patient positioning in the scanner. (2) Operator input required for segmentation. Let $S^{\mathcal{X}}_1, S^{\mathcal{X}}_2, \ldots, S^{\mathcal{X}}_n$ be n sets of scenes which represent n repeat scans, registered and redigitized, of the same subjects and for the same application domain $\mathcal{X}$. In other words, we think of $S^{\mathcal{X}}_1, S^{\mathcal{X}}_2, \ldots, S^{\mathcal{X}}_n$ to represent repeat scans corresponding to $S^{\mathcal{X}}$, with $S^{\mathcal{X}}_1 = S^{\mathcal{X}}$. Let $H_1, H_2, \ldots, H_m$ be m human operators and let M be a particular segmentation method. Let $\mathcal{C}_{O_1}$ and $\mathcal{C}_{O_2}$ be fuzzy segmentations of the same object $\mathcal{O}$ pertaining to the same subject in two repeated trials. $\mathcal{C}_{O_1}$ and $\mathcal{C}_{O_2}$ may have resulted from one of the following situations.

T_1: The same operator segments the same object in the same scene twice by using method M (intra-operator).

T_2: Two operators segment the same object in the same scene once by using method M (inter-operator).

T_3: The same operator segments the same object once in two corresponding scenes in $S^{\mathcal{X}}_i$ and $S^{\mathcal{X}}_j$ $(i \neq j)$ by using method M (inter-scan).

For the given method of segmentation M, all possible pairs $(\mathcal{C}_{O_1}, \mathcal{C}_{O_2})$ for T_1 will allow us to assess *intra-operator* precision of M. Analogously, T_2 and T_3 correspond to the assessment of *inter-operator* and *repeat-scan* (inter-scan) precision. A measure of precision for method M in a trial that produced fuzzy segmentations $\mathcal{C}_{O_1}$ and $\mathcal{C}_{O_2}$ for situation T_i is given by

$$PR^M_{T_i}(\mathcal{O}) = \frac{|\mathcal{C}_{O_1} \cap \mathcal{C}_{O_2}|}{|\mathcal{C}_{O_1} \cup \mathcal{C}_{O_2}|}. \tag{4}$$

$PR^M_{T_i}(\mathcal{O})$ represents the total amount of the tissue that is common to both $\mathcal{C}_{O_1}$ and $\mathcal{C}_{O_2}$ as a fraction of the total amount of tissue in the union of $\mathcal{C}_{O_1}$ and $\mathcal{C}_{O_2}$. $PR^M_{T_i}(\mathcal{O})$ values estimated over the scenes in $S^{\mathcal{X}}_1, S^{\mathcal{X}}_2, \ldots, S^{\mathcal{X}}_n$ utilizing operators $H_1, H_2, \ldots, H_m$ characterize the intra-operator, inter-operator, and repeat-scan (inter-scan) repeatability (respectively for $i = 1,2,3$) of method M. The precision of method M for a given situation $(i = 1,2,3)$ can be characterized by computing the coefficient of variation or confidence intervals of the $PR^M_{T_i}$ values. The precision of any two segmentation methods M_1 and M_2 for each T_i can be compared by comparing the set of $PR^M_{T_i}$ values by using a paired t-test.

Note that just how much the volumes of $\mathcal{C}_{O_1}$ and $\mathcal{C}_{O_2}$ agree (especially for T_1 and T_2) will not constitute a robust measure of precision as illustrated in Fig. 4. This is because $\mathcal{C}_{O_1}$ and $\mathcal{C}_{O_1}$ may have identical volumes but may constitute substantially different delineations. However, situation T_3 is quite different from T_1 and T_2 in that it involves a registration and subsequent interpolation. As demonstrated in [23], because of the errors associated with the latter processes, especially when the object has thin and subtle features (as in peripheral cerebrospinal fluid in the brain), the overlap measure of Eq. (5) may indicate poor precision

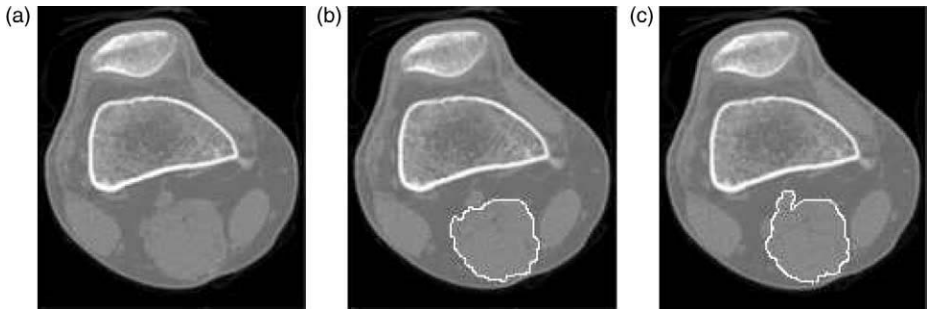

Fig. 4. Segmented objects (muscles) obtained in two different situations T_i and T_j (b) and (c) in a slice of a CT scene of a knee (a). The two segmentations have nearly identical volumes, still they differ substantially.

even when the segmentation is of excellent quality in the repeat scans. To avoid this problem, we suggest that, for T_3, we must use simply the volume of the object computed from the segmentation of original repeat scans (without subjecting repeat acquisitions to registration and then interpolation). In this case, $PR^M_{T_3}(\mathcal{O})$ is defined by

$$PR^M_{T_3}(\mathcal{O}) = 1 - \frac{||\mathcal{C}_{O_1}| - |\mathcal{C}_{O_2}||}{\frac{1}{2}\left[|\mathcal{C}_{O_1}| + |\mathcal{C}_{O_2}|\right]}. \quad (5)$$

A fourth metric $PR^M_{T_4}$ becomes necessary in large clinical trials wherein different sites (possibly with different brands of scanners) are utilized for image acquisition. $PR^M_{T_4}$ will then constitute inter-site precision, and its treatment should be similar to that of $PR^M_{T_3}$. In this case, $S^{\mathcal{X}}_1, S^{\mathcal{X}}_2, \ldots, S^{\mathcal{X}}_l$ will represent l sets of scenes obtained for the same subjects at l different sites for $\mathcal{X}$. If $\mathcal{X}$ involves image acquisition on only one scanner, then there is no need for assessing $PR^M_{T_4}$. We note that for estimating the precision of any method M, surrogates of true delineations and of recognition are not needed. In summary, there are four precision metrics we utilize: $PR^M_{T_1}$, $PR^M_{T_2}$, $PR^M_{T_3}$, and $PR^M_{T_4}$.

2.4.2. Assessment of accuracy

We consider accuracy measures separately for object delineation and recognition.

(a) Delineation. Let $S^{\mathcal{X}}_{td}$ be the set of scenes containing 'true' delineations for the scenes in $S^{\mathcal{X}}$. For any scene $\mathcal{C} = (C, f) \in S^{\mathcal{X}}$, let $\mathcal{C}^M_d$ be the fuzzy segmentation of an object $\mathcal{O}$ of B in $\mathcal{C}$ obtained by using any method M, and let $\mathcal{C}_{td} \in S^{\mathcal{X}}_{td}$ be the corresponding scene of 'true' delineation, all under the application domain $\mathcal{X}$. Let U_d be a subset of C such that it constitutes a *reference superset* with respect to which all delineated regions (true as well as false) within $\mathcal{C}$ can be expressed as a fraction. Let $\mathcal{U}_d$ be the binary scene representing U_d, that is, a scene with domain C and with a scene intensity value of one for all voxels in U_d and a value of 0 for voxels in $C - U_d$. We shall comment on the choice of U_d later on. The only theoretical requirement on U_d is that any delineated region within $\mathcal{C}$ corresponding to $\mathcal{O}$ by any segmentation method be a subset of U_d. Let $\mathcal{C}_{FN} = \mathcal{C}_{td} - \mathcal{C}^M_d$, $\mathcal{C}_{FP} = \mathcal{C}^M_d - \mathcal{C}_{td}$, $\mathcal{C}_{TP} = \mathcal{C}^M_d \cap \mathcal{C}_{td}$, and $\mathcal{C}_{TN} = \mathcal{U}_d - \mathcal{C}^M_d - \mathcal{C}_{td}$, where the operations between scenes are as defined in Eqs. (1)–(3). The following measures are defined to characterize the delineation accuracy of method M under $\mathcal{X}$.

True positive volume fraction,

$$TPVF^M_d(\mathcal{O}) = \frac{|\mathcal{C}_{TP}|}{|\mathcal{C}_{td}|}, \quad (6)$$

True negative volume fraction,

$$TNVF^M_d(\mathcal{O}) = \frac{|\mathcal{C}_{TN}|}{|\mathcal{U}_d - \mathcal{C}_{td}|}, \quad (7)$$

False positive volume fraction,

$$FPVF^M_d(\mathcal{O}) = \frac{|\mathcal{C}_{FP}|}{|\mathcal{U}_d - \mathcal{C}_{td}|}, \quad (8)$$

False negative volume fraction,

$$FNVF^M_d(\mathcal{O}) = \frac{|\mathcal{C}_{FN}|}{|\mathcal{C}_{td}|}. \quad (9)$$

The meaning of these measures is illustrated in Fig. 5 for the binary case. $TPVF^M_d$ indicates the fraction of the total amount of tissue in the true delineation $\mathcal{C}_{td}$ that was covered by method M. $TNVF^M_d$ describes the fraction of the total amount of tissue in the reference region $\mathcal{U}_d$ that is truly not in the object that was also excluded by method M. $FPVF^M_d$ denotes the amount of tissue falsely identified by method M as a fraction of the amount of tissue in $\mathcal{U}_d$ that is truly not in the object. And $FNVF^M_d$ expresses the fraction of tissue in the true delineation $\mathcal{C}_{td}$ that was missed by method M. The following desirable relationships among the above measures can be easily established from Eqs. (1) and (6)–(9).

$$\mathcal{U}_d = \mathcal{C}_{TP} \cup \mathcal{C}_{TN} \cup \mathcal{C}_{FP} \cup \mathcal{C}_{FN}, \quad (10)$$

$$FPVF^M_d(\mathcal{O}) = 1 - TNVF^M_d(\mathcal{O}), \quad (11)$$

$$FNVF^M_d(\mathcal{O}) = 1 - TPVF^M_d(\mathcal{O}). \quad (12)$$

We note that, in view of Eqs. (10)–(12), only two of the four measures are independent. Consequently, only two measures (such as $FPVF^M_d$ and $FNVF^M_d$, or $TPVF^M_d$ and $FPVF^M_d$) need to be specified to describe the delineation accuracy of method M. The above measures are borrowed from statistical decision theory as applied to observer studies [24] but appropriately modified to our situation of segmentation delineation. Continuing along these lines, we define *delineation sensitivity* of method M to be given by $TPVF^M_d(\mathcal{O})$ and *delineation specificity* to be described by $1 - FPVF^M_d(\mathcal{O})$. Clearly, the greater both these entities are, the better is the delineation accuracy of method M.

Some comments are in order regarding the definition in Eqs. (6)–(9) and the choice of U_d. We argue that any alternative definition should satisfy Eqs. (10)–(12). Most

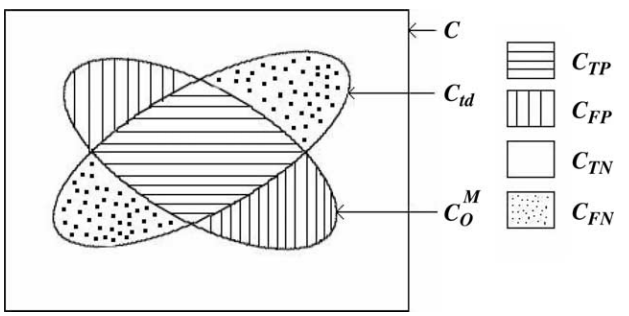

Fig. 5. Illustration of the accuracy factors for delineation for a binary case. Here, $\mathcal{U}_d$ is assumed to be a binary scene with all voxels in the scene domain C set to have a value 1.

likely there cannot be alternatives for the numerators in Eqs. (6)–(9). (Other measures such as those based on the boundary of O have been used in place of the region-based measures described above. Our comments here are applicable to the region-based measures. The boundary-based measures will also have to be, and can be, cast in terms of TP, TN, FP, and FN for appropriately characterizing delineation accuracy). However, for the denominator, other choices seem plausible; for example, $|\mathcal{C}_{TP} \cup \mathcal{C}_{FP} \cup \mathcal{C}_{FN}|$ in Eqs. (6), (8), and (9). Such a choice has been utilized in the past for $TPVF$ (for example in [25]) and for $TPVF$, $FPVF$, and $FNVF$ [26,27]. Another alternative is to use $|\mathcal{C}_{td}|$ in the denominator of all of Eqs. (6)–(9). All such choices lead to the situation of not satisfying Eqs. (10)–(12), unlike the definitions in Eqs. (6)–(9). The satisfaction of Eqs. (10)–(12) is essential for these measures to make sense.

Coming now to the choice of U_d (and hence $\mathcal{U}_d$), one obvious possibility is to take $\mathcal{U}_d = \mathcal{C}$ as in Fig. 5. This has the undesirable property that $TNVF_d^M$ and $FPVF_d^M$ will depend on the size of the scene domain chosen; by making C large, these two factors can be changed arbitrarily. Another possible choice for $\mathcal{U}_d$ is the body region B as it manifests in scene $\mathcal{C}$. In (macroscopic) medical imaging, this usually corresponds to the foreground region of the scene. In comparing methods based on the same scene data sets, these choices may not matter much. These issues obviously require further research and deliberation. We argue that a standard evaluation framework (with the five components (F1)–(F5) mentioned earlier) is essential to carry out meaningful and exchangeable segmentation performance evaluation measures.

Fig. 6 presents an example showing the above four factors for the application domain of brain parenchymal volume estimation via MRI T2 and PD scenes and by using the fuzzy connectedness segmentation method [23].

(b) Recognition. At present, the existing segmentation algorithms seem to focus mainly on delineation without considering in their design the ability to capture salient feature/landmark information. It is conceivable, however, that methods can be devised to recognize important landmarks as part of the segmentation process. In such a case, we may formulate measures for recognition exactly along the lines described for delineation. If $\mathcal{C}_r^M$ is the scene representing the landmarks recognized by method M, then $TPVF_r^M$, $TNVF_r^M$, $FPVF_r^M$, and $FNVF_r^M$ can be defined exactly as in Eqs. (6)–(9) once an appropriate choice is made for the reference superset U_r for recognition (and for the corresponding scene $\mathcal{U}_r$). At the current state of affairs, such algorithms for recognition as part of the segmentation process do not exist. Therefore, for now we suggest using the following two measures for recognition, utilizing the result of delineation $\mathcal{C}_d^M$ to judge the ability of M to capture important landmark information.

$$TPVF_r^M(\mathcal{O}) = \frac{|\mathcal{C}_{tr} \bullet \mathcal{C}_d^M|}{|\mathcal{C}_{tr}|}, \tag{13}$$

$$FNVF_r^M(\mathcal{O}) = \frac{|\mathcal{C}_{tr} \circ \bar{\mathcal{C}}_d^M|}{|\mathcal{C}_{tr}|}, \tag{14}$$

where $\bullet$ and $\circ$ are as defined in Eqs. (2) and (3). The idea here is to determine what portion of the spread region of the landmarks/features is captured by $\mathcal{C}_d^M$. The total weight in this captured region as a fraction of the total weight in $\mathcal{C}_{tr}$ defines $TPVF_r^M$ for characterizing the accuracy of the qualitative (recognition) aspect of segmentation by method M. Analogously, $FNVF_r^M$ specifies the fraction of the total weight in $\mathcal{C}_{tr}$ that is missed by method M. Fig. 7 illustrates these ideas for the application domain of determining the kinematics of the ankle joint complex via MRI [28]. Here we focused on the problem of segmenting one of the bones of the joint, namely the talus ($\mathcal{O}$). Two experts (BEH, JW) compiled a set of five features which included the following: (i) The superior surface (articular surface) of the body of the talus. (ii) The inferior surface (posterior articular facet) of the body of the talus. (iii) The talus head. (iv) The middle calcaneal articular surface on the inferior surface of the neck of the talus. (v) A distinct point defining the posterolateral edge of the sinus tarsi. We have assigned equal weighting for all features. Subsequently $S_{tr}^{\mathcal{X}}$ was created from the repeated (three times) delineation provided by them for these features on a set of five scenes. Fig. 7a shows a slice of one of the scenes $\mathcal{C}$ in $S^{\mathcal{X}}$, and Fig. 7b shows the appearance of three of these features (i)–(iii) on the corresponding slice.

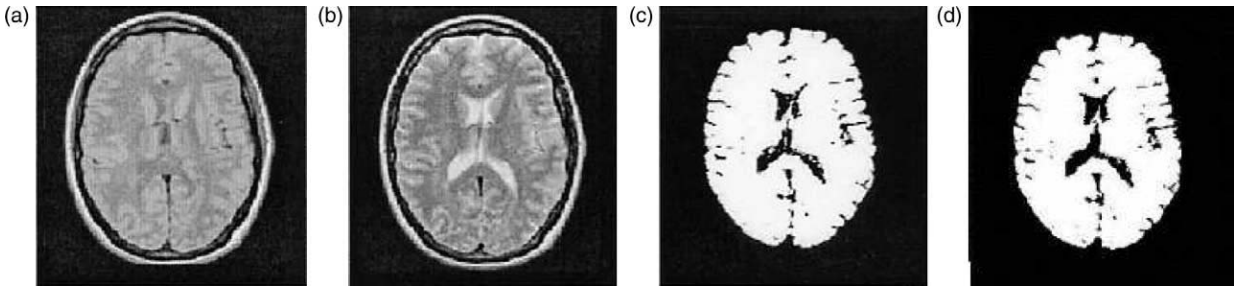

Fig. 6. Assessment of accuracy of segmentation. (a) and (b) A slice of a PD and T2 MRI scene of a patient's brain. (c) The result of fuzzy connectedness segmentation of brain parenchyma (in 3D). (d) 'True' delineation obtained by manual correction of the fuzzy connectedness segmentation. For this example, $FNVF = 1.9\%$, $FPVF = 0.2\%$, $TPVF = 98.1\%$, $TNVF = 99.8\%$.

Fig. 7. (a) One slice of the human ankle MRI scene. (b) The corresponding slice of C_{tr}. (c) The slice of C_{tr} overlayed on the corresponding slice of C_d^M obtained by using the live-wire method.

Fig. 7c displays the corresponding slice of C_{tr} overlayed on the corresponding slice of C_d^M. M in this case is the live-wire method [26].

In summary, there are six accuracy metrics that this framework will utilize only three of which are independent measures: $FNVF_d^M$, $FPVF_d^M$, $TNVF_d^M$, $TPVF_d^M$, $FNVF_r^M$, and $TPVF_r^M$.

2.4.3. Assessment of efficiency

We note that for any method M, the precision and accuracy metrics influence one another in a complex manner in the following sense. An attempt to improve accuracy is usually accompanied by a compromise in precision and vice versa. As an example, consider M to represent the method of thresholding based on a fixed threshold value, as illustrated in Fig. 8, where $\langle T,B,P\rangle$ is the application domain considered in Fig. 6. Obviously $PR_{T_1}^M$ and $PR_{T_2}^M$ are both one. However, with repeat scan (Fig. 8a and b) there is much variation in the result (Fig. 8c and d) and $PR_{T_3}^M$ becomes 0.702. The 'true' delineations for the two scans of Fig. 8a and b are shown in Fig. 8e and f, respectively. It is clear that, although this method has high precision (except for the third factor $PR_{T_3}^M$) and degree of automation (efficiency), its accuracy is poor: $FNVF_d^M = 0.142$, and $FPVF_d^M = 0.10$. A possible way of improving accuracy of M is to modify M by having a human operator correct the results post-hoc. This will of course bring down both efficiency and precision. What most segmentation methods strive for is to try to have as few free parameters as possible and then to juggle among these three groups of factors (precision, accuracy, and efficiency) in setting up optimal values for the parameters. Some segmentation methods require other forms of per-scene human help also, such as for initialization (seed specification, initial boundary specification) and for any per-scene algorithm training needed. We denote the total human time

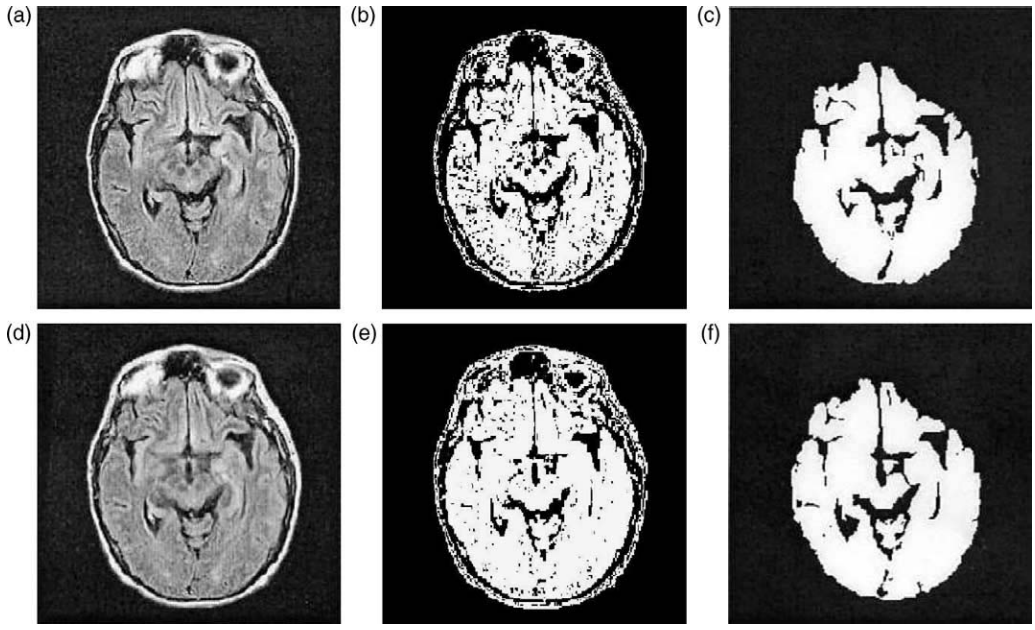

Fig. 8. (a) and (b) Two corresponding slices after registration of a pair of repeat scans (with a short time gap in between scans) of a patient's brain. (c) and (d) Segmentation of (a) and (b) by fixed thresholding. The object of interest is brain parenchyma. (e) and (f) 'True' segmentation of (a) and (b) obtained as for Fig. 6.

required by M in this manner for each scene by $t^M_{h_1}$. Another human time component needed, denoted $t^M_{h_2}$, is for any one-time (and not per-scene) algorithm training needed.

In addition to the human help required, and analogous to the two components $t^M_{h_1}$ and $t^M_{h_2}$, there are two components of computational time required by M, denoted by $t^M_{c_1}$ and $t^M_{c_2}$. They represent, respectively, the computation time required for segmenting each scene and the computation time required for one-time (and not per-scene) algorithm training. We note that $t^M_{h_1}$ is the most crucial among these four time factors.

Efficiency of method M refers to its practical viability in terms of the above four factors. Most methods require some human help and the claim of 'totally automatic' for a method M in an application domain $\langle T,B,P\rangle$ is not valid unless perfect (or high) precision and accuracy is demonstrated for M in $\langle T,B,P\rangle$ over a large (essentially infinite) number of scenes. It is clear therefore, that precision, accuracy, and efficiency factors have a complex interdependency, and the measurement of the efficiency factors is practically highly relevant to distinguish among methods that have otherwise comparable precision and accuracy but vastly differing efficiency factors, particularly $t^M_{h_1}$ and $t^M_{h_2}$. A sensible way of combining the efficiency factors is via the dollar cost incurred. Assuming a trained technician to be a standard human operator H, H's salary γ will determine the cost per time unit, and, hence, the weight to be given to $t^M_{h_1}$ and $t^M_{h_2}$. Similarly, the cost (λ) per unit computer time will determine the weight to be given to $t^M_{c_1}$ and $t^M_{c_2}$. Obviously γ and λ can be determined and fixed within an evaluation framework, although they will have to be updated on a regular basis. For a given $\mathcal{X}$, the overall efficiency E^M of method M can be given by $E^M = g(t^M_{h_1},t^M_{h_2},t^M_{c_1},t^M_{c_2})$, where g is a function that converts time factors into dollar cost for the total cost incurred in segmenting each scene in the set of scenes $S^{\mathcal{X}}$ by utilizing γ and λ. In summary, this framework will utilize five efficiency metrics: $t^M_{h_1}$, $t^M_{h_2}$, $t^M_{c_1}$, $t^M_{c_2}$, and E^M. For comparing methods (Section 3), either the efficiency factor or directly the time factors can be utilized.

3. How to compare methods

The procedure for comparing two methods M_1 and M_2 under a given $\langle T,B,P\rangle$ consists of the following steps.

(1) Collect sets of scenes $S^{\mathcal{X}}_1,S^{\mathcal{X}}_2,\ldots,S^{\mathcal{X}}_n$ corresponding to n repeat scans of the scenes in $S^{\mathcal{X}}$ acquired for $\langle T,B,P\rangle$. Produce scenes $S^{\mathcal{X}}_{td}$ and $S^{\mathcal{X}}_{tr}$ representing surrogate of true delineation and of recognition for the scenes in $S^{\mathcal{X}}$.
(2) Optimize the implementations of M_1 and M_2 for $\langle T,B,P\rangle$. For methods M_1 and M_2, have operators $H_1,H_2,\ldots,H_m$ repeat segmentations of scenes in $S^{\mathcal{X}}$. Have one operator segment scenes in $S^{\mathcal{X}}_1,S^{\mathcal{X}}_2,\ldots,S^{\mathcal{X}}_n$ for methods M_1 and M_2.
(3) For $i=1,2,3,4$, determine all possible values of $PR^{M_1}_{T_i}$ and $PR^{M_2}_{T_i}$.
(4) Knowing $S^{\mathcal{X}}_{td}$ and $S^{\mathcal{X}}_{tr}$ and the segmentations of the scenes in $S^{\mathcal{X}}$ produced by M_1 and M_2, compute $FNVF^{M_j}_d$, $FPVF^{M_j}_d$, $TPVF^{M_j}_d$, $TNVF^{M_j}_d$, $TPVF^{M_j}_r$, and $FNVF^{M_j}_r$, for $j=1, 2$.
(5) Record $t^{M_j}_{c_1},t^{M_j}_{c_2},t^{M_j}_{h_1},t^{M_j}_{h_2}$ for $j=1, 2$ during the segmentation experiments, and from these compute the efficiency metric E^{M_j}.
(6) For each method M_j, we get a set of values for each of the 15 parameters: $PR^{M_j}_{T_1}$, $PR^{M_j}_{T_2}$, $PR^{M_j}_{T_3}$, $PR^{M_j}_{T_4}$, $FNVF^{M_j}_d$, $FPVF^{M_j}_d$, $TPVF^{M_j}_d$, $TNVF^{M_j}_d$, $TPVF^{M_j}_r$, $FNVF^{M_j}_r$, $t^{M_j}_{h_1}$, $t^{M_j}_{h_2}$, $t^{M_j}_{c_1}$, $t^{M_j}_{c_2}$, and E^{M_j}.

Considering only two independent accuracy parameters and four $\left(t^{M_j}_{h_1}, t^{M_j}_{h_2}, t^{M_j}_{c_1}, t^{M_j}_{c_2}\right)$ (or one (E^{M_j})) among the efficiency parameters, we get a set of 12 (or 9) parameters altogether. There are several choices for the statistical analysis of these 12 (or 9) sets of values.

(a) Do a paired t-test of the two sets of values for each parameter for the two methods.
(b) Combine the 12 (or 9) parameters for each method M_j by a weighted sum, the weight reflecting the importance given to that parameter for $\langle T,B,P\rangle$ and then do a paired t-test of the resulting single parameter.
(c) Do a multivariate analysis of variance [29] considering all 12 (or 9) parameters to determine if there is a statistically significant difference in performance between methods M_1 and M_2.

As an example, we display in Table 1 8 of the 12 parameters for the application domain illustrated in Fig. 7. The two methods compared are M_1 = fuzzy connectedness (*FC*) [23], and M_2 = live-wire (*LW*) [26] which is a user-steered boundary segmentation method and the steering was provided by a non-expert in the application domain (YZ). $S^{\mathcal{X}}$ consisted of a set of 20 MRI scenes. Because of the connective tissues (ligaments and tendons) and because

Table 1
Some of the metrics computed from 20 scenes for the two methods *FC* and *LW*

	$PR^M_{T_1}$	$FNVF^M_d$	$FPVF^M_d$	$TPVF^M_r$	$t^M_{h_1}$ (min)	$t^M_{h_2}$ (min)	$t^M_{c_1}$ (min)	$t^M_{c_2}$ (min)
FC	0.99 ± 0.015	0.13 ± 0.043	0.04 ± 0.018	0.78 ± 0.035	1.9 ± 01.3	1	1.85 ± 0.24	0
LW	0.96 ± 0.011	0.03 ± 0.010	0.03 ± 0.017	0.94 ± 0.018	8.5 ± 1.3	1	~0	0
Comparison	$FC>LW$ ($p\leq0.05$)	$LW>FC$ ($p\leq0.05$)	–	$LW>FC$ ($p\leq0.05$)	$FC>LW$ ($p\leq0.05$)	–	$LW>FC$ ($p\leq0.05$)	–

cortical bone yields very little MR signal, this is a difficult segmentation problem. $S_{td}^{\mathcal{X}}$ has been previously created for this application domain by well trained students of podiatric medicine via *LW* and subsequently scrutinized (and corrected if necessary) by an expert (BEH). The two numbers in the table represent the mean and the standard deviation of the metric over $S^{\mathcal{X}}$. The *p*-value for a paired *t*-test on each of the metrics comparing the two methods is also listed when the difference is statistically significant. We note from the table that, although *FC* is more efficient from the consideration of operator help needed ($t_{h_1}^M$) and more precise ($PR_{T_1}^M$), it is less accurate (in both delineation and recognition) and requires more computational time than *LW* for this application domain ('$>$' indicates better than). (For *LW*, the live-wire segments are computed and displayed in real time [26], and therefore, $t_{c_1}^M$ is negligible). Clearly, the preference for the methods in an application domain depends very much on which of these metrics is crucial for that domain.

4. Concluding remarks

(1) If the surrogates of truth are highly reliable (gold standard), then it may appear that there is no need to evaluate precision, and accuracy analysis would be sufficient. However, accuracy analysis will then have to consider intra-operator, inter-operator, repeat-scan, and inter-site variations. We feel that it is best to relegate the analysis of these variations to a separate group, namely precision. Therefore, the factors describing precision, accuracy, and efficiency are all essential in assessing the performance of segmentation methods.

(2) A descriptive answer in terms of the various parameters gives a more meaningful and complete assessment of the methods than an answer to the overall question 'Is method M_1 better than M_2 under the application domain?'

(3) Since, most segmentation methods in practice consider only delineation, we suggest that, at a minimum, the following set of seven parameters should be evaluated: $PR_{T_1}^M$, $PR_{T_2}^M$, $PR_{T_3}^M$, $FNVF_d^M$, $FPVF_d^M$, $t_{h_1}^M$, $t_{c_1}^M$ for any given method M.

(4) General statements about the merit of segmentation algorithms cannot be made independent of the application domain $\langle T,B,P\rangle$. The evaluative results of two methods M_1 and M_2 observed under one $\langle T,B,P\rangle$ may not foretell anything about their comparative behavior for a different $\langle T,B,P\rangle$.

(5) We have proposed a method to incorporate into the evaluation method the aspect of how well key features (landmarks) of an object that are considered important for $\langle T,B,P\rangle$ are captured in the segmentation. We are able to include this qualitative aspect of recognition also within the same common framework of evaluation.

(6) The four components of efficiency are essential, $t_{h_1}^M$ being the most crucial among these. There is no such thing as 'an automatic segmentation method'. Any method may fail (for example, it may produce high $FNVF_d^M$ and/or $FPVF_d^M$ for a particular data set) if a sufficiently large set of scenes is processed, and then it will need human intervention. 'Automatic' is only a design intent and not necessarily the end result for a segmentation method. Therefore, the phrase has no meaning, even for a particular application domain, unless the method's efficiency is proven to be 100% (for all four factors) over a large (essentially infinite) number of data sets with acceptable precision and accuracy in the application domain.

(7) The factors describing precision, accuracy, and efficiency are interdependent. To simultaneously improve all three factors for a method is usually difficult and requires considerable research. An attempt to increase accuracy may be accompanied by a decrease in efficiency and/or precision.

(8) Once the surrogates are determined, the framework can be easily implemented and utilized to evaluate any image segmentation methods.

(9) A framework with the five components (F1)–(F5) mentioned in Section 1 becomes essential to carry out meaningful, exchangeable, and widely accepted segmentation performance evaluation. Further work is needed in all of these components. Further research is also needed in the definition of accuracy measures. We have focused on region-based measures in this paper. Boundary-based or other (even hybrid) strategies may be more relevant in certain application domains. (For example, when the object shape is such that its surface area to volume ratio is high—that is, when the number of voxels in the boundary approaches the number of voxels constituting the region occupied by the object—small changes in segmentation would yield large changes in the precision and accuracy measures. For such objects, perhaps boundary-based measures are more appropriate). Their definition satisfying conditions similar to those in Eqs. (10)–(12) requires further work. Most segmentation algorithms behave like human observers in that their performance cannot be completely characterized by measures derived at one operating point, which is decided by the values assigned to the parameters of the method. There is even no systematic approach available for setting the values of the parameters of segmentation methods optimally. Methods akin to ROC analysis are needed to more completely characterize the *range of behavior* of the accuracy of segmentation methods.

(10) Some comments are in order regarding the terms 'evaluation' vis-a-vis 'validation'. It is wrong to call the process of evaluating a segmentation method M in an application domain $\mathcal{X}$ a 'validation process'. Since, the level of performance of M in $\mathcal{X}$ is unknown, a neutral

term like 'evaluation' is more appropriate to describe the process. Only when the evaluation is completed and if the level of performance of M in $\mathcal{X}$ becomes acceptable, it is appropriate to describe M as being validated in $\mathcal{X}$. In the phrase 'validation process', there is a hint of presumption and wishful thinking. We, therefore, suggest that *evaluation* be used to describe this process.

Acknowledgements

The research reported here is supported by DHHS grants NS 37172 and AR46902. The authors are grateful to Gul Moonis for Fig. 1, Punam Saha for Figs. 2 and 3, and to Laszlo Nyúl for Figs. 6 and 8.

References

[1] Pal NR, Pal SK. A review on image segmentation techniques. Pattern Recogn 1993;26:1277–94.
[2] Pham D, Xu C, Prince J. Current methods in medical image segmentation. Annu Rev Biomed Eng 2000;2:315–37.
[3] Yasnoff W, Mui J, Bacus J. Error measures for scene segmentation. Pattern Recogn 1977;9:217–31.
[4] Chalana V, Kim Y. A methodology for evaluation of boundary detection algorithms on medical images. IEEE Trans Med Imaging 1997;16:642–52.
[5] Hoover A, Jean-Baptiste G, Jiang X, Flynn P, Bunke H, Goldgof D, et al. An experimental comparison of range image segmentation algorithms. IEEE Trans Pattern Anal Mach Intell 1996;18:673–89.
[6] Zhang Y. A survey on evaluation methods for image segmentation. Pattern Recogn 1996;29:1335–46.
[7] Bowyer K. Validation of medical image analysis techniques. In: Beutel J, Kundel H, Metter Rvan, editors. The Handbook of Medical Imaging. SPIE; 2000.
[8] Loughlin M, Carlbom I, Busch C, Douglas T, Egevad L, Frimmel H, et al. Three-dimensional modeling of biopsy protocols for localized prostate cancer. Comput Med Imaging Graph 1998;22:229–38.
[9] McFarland E, Brink J, Pilgram T, Heiken J, Balfe D, Hirselj D, et al. Spiral CT colonography: reader agreement and diagnostic performance with two- and three-dimensional image-display techniques. Radiology 2001;218:375–83.
[10] De Graaf C, Koster A, Vinken K, Viergever M. A methodology for the validation of image segmentation methods. IEEE Symp Comput Based Med Syst Proc 1992;17–24.
[11] Dawant B, Hartmann S, Thirion J, Maes F, Vandereulen D, Demaerel P. Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformation. Part 1, methodology and validation on normal subjects. IEEE Trans Med Imaging 1999;10:909–16.
[12] Brown M, Feng W, Hall T, McNitt-Gray M, Churchill B. Knowledge-based segmentation of pediatric kidneys in CT for measurement of parenchymal volume. J Comput Assist Tomogr 2001;25:639–48.
[13] Huttenlocher D, Klanderman G, Rucklidge W. Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 1993;15:850–63.
[14] Udupa JK, LeBlanc VR, Schmidt H, Imielinska C, Saha PK, Grevera GJ, et al. A methodology for evaluating image segmentation algorithms. SPIE Proc 2002;4684:266–77.
[15] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004; 23(7):903–21.
[16] Filippi M, Horsfield M, Bressi S, Martinelli V, Baratti C, Reganatr P. Intra- and inter-observer variability of brain MRI lesion volume measurements in multiple sclerosis: a comparison of techniques. Brain 1995;118:1593–600.
[17] Wicks D, Tofts P, Miller D, Du Boulay G, Feinstein A, Sacares R. Volume measurement of multiple sclerosis lesions with magnetic resonance images: a preliminary study. Neuroradiology 1992;34: 475–9.
[18] Patty D, Li D. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis, II. MRI analysis results of a multicenter, randomized, double blind, placebo-controlled trial,UBC MS/MRI Study Group and the IFNB Multiple Sclerosis Study Group. Neurology 1993;43:662–7.
[19] Pizer SM. Intensity mappings to linearize display devices. Comput Graph Image Proc 1981;17:262–8.
[20] Nyúl LG, Udupa JK. On standardizing the MR image intensity scale. Magn Reson Med 1999;42:1072–81.
[21] Browne JA, Herman GT, Odhner D. SNARK93—a programming systems for image reconstruction from projections. Department of Radiology, University of Pennsylvania, Technical report MIPG198; 1993.
[22] Collins DL, Zijdenbos AP, Kollokian V, Sled JG, Kabani NJ, Holmes CJ, et al. Design and construction of a realistic digital brain phantom. IEEE Trans Med Imaging 1998;17:463–8.
[23] Nyúl LG, Udupa JK. A Protocol-independent brain MRI segmentation method. SPIE Proc 2002;4684:1588–99.
[24] Metz CE, Goodenough DJ, Rossmann K. Evaluation of receiver operating characteristic curve data in terms of information theory with applications in radiology. Radiology 1973;108:297–303.
[25] van Ginneken B, Frangi AF, Staal JJ, ter Haar Romeny BM, Viergever MA. Active shape model segmentation with optimal features. IEEE Trans Med Imaging 2002;21:924–33.
[26] Falcão AX, Udupa JK, Samarasekera S, Sharma S. User-steered image segmentation paradigms: live wire and live lane. Graph Models Image Process 1998;60:233–60.
[27] Udupa JK, Wei L, Samarasekera S, Miki Y, Buchem MA, Grossman RI. Multiple sclerosis lesion quantification using fuzzy connectedness principles. IEEE Trans Med Imaging 1997;16: 598–609.
[28] Stindel E, Udupa JK, Hirsch BE, Odhner D. An in vivo analysis of the peri-talar joint complex based on MR imaging. IEEE Trans Biomed Eng 2001;48:236–47.
[29] Rencher AC. Multivariate analysis of variance. In: Methods of multivariate analysis. New York, NY: Wiley; 1995. p. 174–257.