A Methodology for Evaluating Image Segmentation Algorithms

Jayaram K. Udupa[a], Vicki R. LaBlanc[b], Hilary Schmidt[b], Celina Imielinska[c], Punam K. Saha[a], George J. Grevera[a], Ying Zhuge[a], Pat Molholt[c], Yinpeng Jin[d], Leanne M. Currie[b]

[a]Medical Image Processing Group - Department of Radiology - University of Pennsylvania
423 Guardian Drive - 4th Floor Blockley Hall - Philadelphia, Pennsylvania - 19104-6021

[b]Columbia University College of Physicians and Surgeons, Office of Scholarly Resources and Center for Education Research and Evaluation - New York, New York

[c]Columbia University College of Physicians and Surgeons, Office of Scholarly Resources and Department of Medical Informatics

[d]Columbia University College of Physicians and Surgeons, Office of Scholarly Resources and Department of Biomedical Engineering

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 (reproducibility), accuracy (agreement with truth), and efficiency (time taken) – need to be considered for both recognition and delineation. To assess precision, we need to choose a figure of merit (FOM), repeat segmentation considering all sources of variation, and determine variations in FOM 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. To assess efficiency, both the computational and the user time required for algorithm and operator training and for algorithm execution should be measured and analyzed. Precision, accuracy, and efficiency are interdependent. It is difficult to improve one factor without affecting others. Segmentation methods must be compared based on all three factors. 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 – the process of defining objects in images – is the most crucial among all operations done on acquired images. Even seemingly unrelated operations such as image (gray level) display [1], interpolation [2], filtering [3], and registration [4] 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 nearly four decades of research [5], segmentation remains a challenging problem in image processing and computer vision.

A related, tightly coupled problem is the evaluation of segmentation methods. Part of the difficulty faced in developing segmentation methods is the lack of a unified framework for their evaluation. Methods published expressly

for addressing segmentation evaluation are rare and are very restricted in scope [6, 7]. Evaluation methods proposed in papers reporting new segmentation algorithms are even more ad hoc and restricted. In spite of the numerous workshops, panel discussions, and special sessions devoted to this subject in many conferences, nothing tangible has resulted.

1.2 Purpose

We have been working on image segmentation since the 1970s [8] and have been thinking about a framework for evaluating segmentation algorithms for the past 7 years. The purpose of this paper is to describe a unified framework for segmentation evaluation that resulted from our investigation on developing a variety of segmentation algorithms [9-15] and their use and evaluation in a variety of medical applications [16-22]. This paper does not actually compare any particular segmentation algorithms but describes the concepts of evaluation with examples.

2. THE METHODOLOGY

2.1 Notation

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.

A: An application or 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 description of a particular algorithm α for a given application domain <A, B, P> that signals high performance for α may tell nothing at all about α for a different application domain <A′, B′, P′>. 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 in B for which images are acquired; example: brain tumor.

Scene: A 3D volume image, denoted by $\mathcal{C} = (C, f)$, where C is a rectangular array of voxels, 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 : A set of scenes acquired for the same given application domain <A, B, P>.

2.2 Segmentation

Segmentation of an object O in a given scene acquired for an application domain <A, B, P> is the process of defining the region/boundary of O in the given scene. It consists of two related tasks – recognition and delineation. *Recognition* is a high-level and qualitative task of determining roughly the whereabouts of the object in the scene. *Delineation* is a lower-level and quantitative task of specifying the precise location and extent of the object's region/boundary in the scene. Knowledgeable humans can outperform computer algorithms in the recognition task, whereas algorithms can be devised that can do delineation better than humans.

We assume that the output of any segmentation algorithm corresponding to a given scene $\mathcal{C} = (C, f)$ is a set $O \subset C$ of voxels. This set represents the region occupied by (the support of) an object $\mathcal{O}$ of B in $\mathcal{C}$. The *fuzzy object defined by* $\mathcal{O}$ *in* $\mathcal{C}$ is a scene $\mathcal{C}_O = (C, f_o)$, where, for any $c \in C$,

$$f_o(c) = \begin{cases} \eta(f(c)), & \text{if } c \in C \\ 0, & \text{otherwise .} \end{cases} \quad (1)$$

We shall (for simplicity) call C_O itself a fuzzy object. Here η is a function that assigns a degree of objectness to every voxel *c* in *O* depending on the scene intensity *f*(*c*). We shall always denote a hard segmentation in *C* of an object ○ in B by *O* and the corresponding fuzzy object by C_O .

The *efficacy* of any segmentation method *M* in an application domain <A, B, P> is to be measured in terms of three factors: *Precision* which represent repeatability of segmentation taking into account all subjective actions required in producing the result; *Accuracy*, which denotes the degree to which the segmentation agrees with truth; *Efficiency*, 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 entire segmentation. Our methodology attempts to capture both recognition and delineation within the same framework in the factors considered for evaluation.

We will use the following operations on fuzzy objects. Let $C_{ox} = (C, f_x)$, $C_{oy} = (C, f_y)$ and $C_{oz} = (C, f_z)$ be any fuzzy objects defined by the same object O in a scene C . Then, the *cardinality* $|C_{ox}|$ of the fuzzy object C_{ox} is defined as $|C_{ox}| = \sum_{c \in C} f_x(c)$. Fuzzy set *union* $C_{oz} = C_{ox} \cup C_{oy}$ is defined by, for any $c \in C$ $f_z(c) = \max(f_x(c), f_y(c))$. Fuzzy set *intersection* $C_{oz} = C_{ox} \cap C_{oy}$ is defined by, for any $c \in C$, $f_z(c) = \min(f_x(c), f_y(c))$.

Fuzzy set difference $C_{oz} = C_{ox} - C_{oy}$ is defined by, for any $c \in C$,

$$f_z(c) = \begin{cases} f_x(c) - f_y(c), & \text{if } f_y(c) \geq 0 \\ 0, & \text{otherwise .} \end{cases} \quad (2)$$

A fuzzy masking operation $C_{oz} = C_{ox} \bullet C_{oy}$, called *inside*, is defined by, for any $c \in C$,

$$f_z(c) = \begin{cases} f_x(c), & \text{if } f_y(c) \neq 0 \\ 0, & \text{otherwise .} \end{cases} \quad (3)$$

Another fuzzy masking operation $C_{oz} = C_{ox} \circ C_{oy}$ called *outside* is defined by, for any $c \in C$,

$$f_z(c) = \begin{cases} f_x(c), & \text{if } f_y(c) = 0 \\ 0, & \text{otherwise .} \end{cases} \quad (4)$$

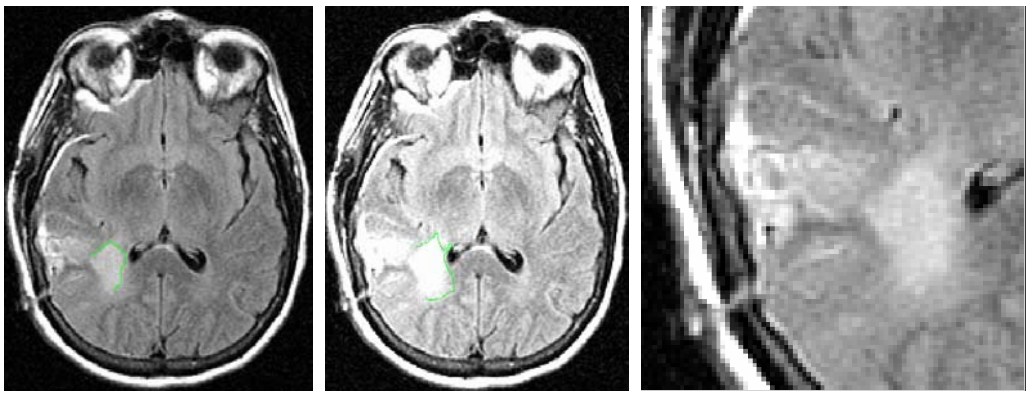
2.3 Surrogate of Truth

For patient images, since it is impossible to establish absolute true segmentation, some surrogate of truth is needed. Our basic premise in developing this framework is that humans outperform computer algorithms in recognition tasks, while computer algorithms are far more efficacious in delineation than humans. Accordingly, the surrogates that are used reflect this premise. We will treat the delineation and recognition aspects separately.

2.3.1 Delineation

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

(1) Manual Delineation: Object boundaries are traced or regions are painted manually by experts (see Figure 1). Corresponding to a given set S of scenes for the application domain <A, B, P>, manual delineation produces a set S_{td} of scenes representing the fuzzy objects defined by the same object represented in the scenes in S. Manual delineation produces a hard set O for each scene $\subset$ in S, which is converted to a fuzzy object via Equation (1). When object regions/boundaries are fuzzy or very complex (fractal like) in a given scene, manual delineation becomes very ill defined. For example, in Figure 1, it is difficult to decide what aspects of the edematous region of the tumor should be included/excluded. Further, to minimize variability, it is important to follow strict protocols for window level and width setting, magnification factor, and the interpolation method used for slice display (Figure 1), and the method of tracing/painting. Multiple repetitions of segmentation by multiple-operators should be performed. There are several ways of averaging the results to get S_{td}. The binary objects (O) segmented in each scene $\subset \in S$ in multiple trials may be averaged first and then the fuzzy object may be computed via Equation (1), or the fuzzy objects C_O may be computed first for the multiple trials which may be averaged. The later is perhaps a better strategy. Manual delineation is inherently binary; that is, it cannot specify tissue percentages. We convert these binary results into fuzzy objects via Equation (1). However, if only binary segmentation is desired, then the manual segmentations are output as binary scenes. In that case, S_{td} contains binary scenes.

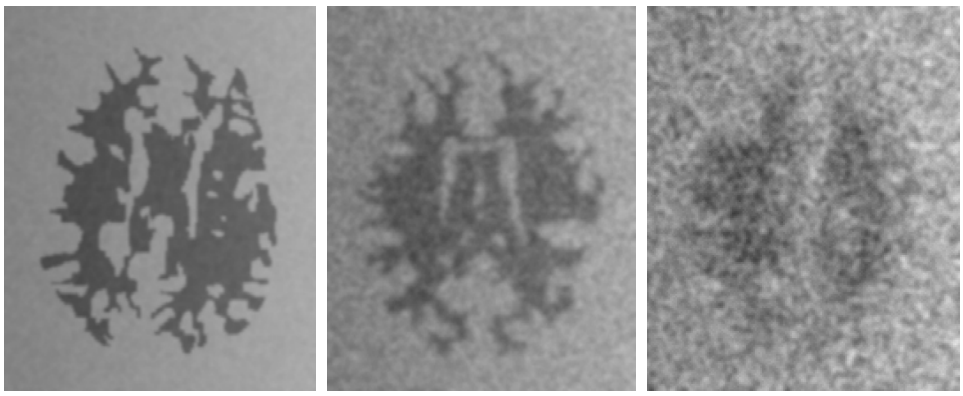


(a) **(b)** **(c)**

FIGURE 1: A slice from the 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.

(2) Mathematical Phantoms: A set of mathematical phantoms is created to depict the application domain <A, B, P> as realistically as possible in terms of blurring, relative tissue contrast and heterogeneity, noise, and background inhomogeneity (see Figure 2) in the scenes. The starting point for this simulation is a set S_{td} of binary scenes (true

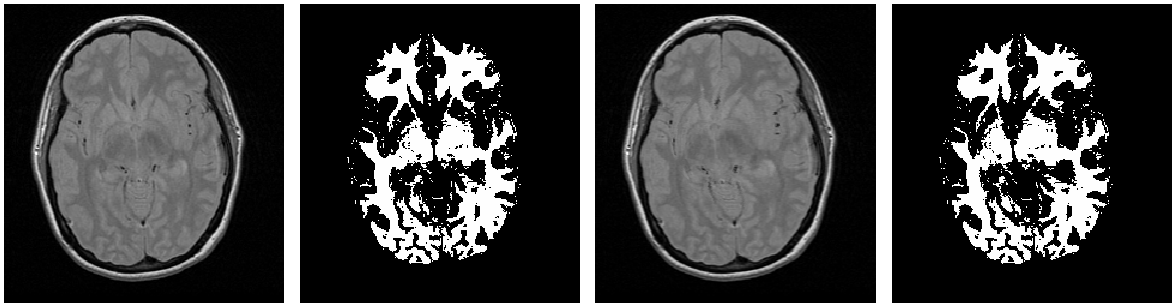
delineation is known to begin with). Each scene in S_{td} is gradually corrupted to yield the actual set of scenes S. We may also start with gray scenes depicting true fuzzy objects and then follow the same procedure.



(a) **(b)** **(c)**

FIGURE 2: White matter (WM) in a gray matter background, simulated by segmenting WM from real MR images and by adding blur, noise, background variation to various degrees: (a) low, (b) medium, and (c) high.

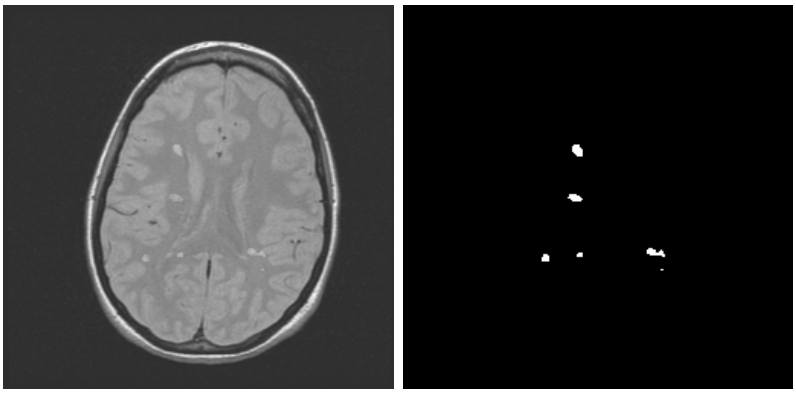
(3) Simulated Scenes I: Use the method of mathematical phantoms described above to generate scenes and apply to both the segmentations and the simulated scenes known 3D deformations (to capture variations that exist among patients) to generate more scenes and their segmentations. The same method is applicable to the method of manual segmentation also (see Figure 3). The complete set of scenes (original + deformed) in this case constitutes S, and the complete set of segmentations represents S_{td}.



(a) **(b)** **(c)** **(d)**

FIGURE 3: Simulating more scenes (c) and their "true" segmentations (d) from existing scenes (a) and their manual segmentation (b) by applying known realistic deformations.

(4) Simulated Scenes II: Another 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 cut-out is segmented carefully by using an appropriate segmentation method. This should not be difficult since the cut-out 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 and S_{td}, respectively. See Figure 4.



(a) **(b)**

FIGURE 4: A slice (a) of a scene simulated from an acquired MR proton density scene of a Multiple Sclerosis patient's brain and its "true" segmentation (b) of the lesions.

2.3.2 Recognition

The approach for ensuring that information related to certain key features or landmarks related to the object (the recognition aspect) is included in the surrogate used for assessing accuracy of segmentation is as follows.

(1) Compile a list of features/landmarks that are vital for <A, B, P> through help from a set of experts (radiologists, surgeons, anatomists).

(2) Each expert assigns a score to each feature to indicate its level of importance in <A, B, P>.

(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].

(4) Have experts locate features in scenes in S repeatedly.

(5) Use the mean location and spread information and the mean vector F to generate a scene C_{tr} (for each scene $C \in S$) which is a composite of the Gaussian weighted scores for all features in the set. In this composite scene $C_{tr} = (C, f_{tr})$, a high value $f_{tr}(c)$ for a voxel $c \in$ C indicates that c is both close to the mean location for a particular feature and the importance of the feature is high. We may also think of generating a scene C_{tr}^{i} for each feature i in F or make C_{tr} a vectorial scene. Alternatively, these individual scenes C_{tr}^{i} may be combined into a composite scene C_{tr} as indicated above by taking an average or a fuzzy union. Fuzzy union is perhaps more appropriate. In any case, let S_{tr} denote the set of resulting scenes containing information about truth in recognition.

2.4 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 $S_1 (= S)$, S_2,….,S_n be n sets of scenes which represent repeat scans, registered and redigitized, of the same subjects and for the same application domain <A, B, P>. Let H_1, H_2,….,H_m be m human operators and let M be a particular segmentation method. Let C_{O1} and C_{O2} be segmentations (fuzzy segmented objects) of the same object O in two repeated trials. C_{O1} and C_{O2} 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).

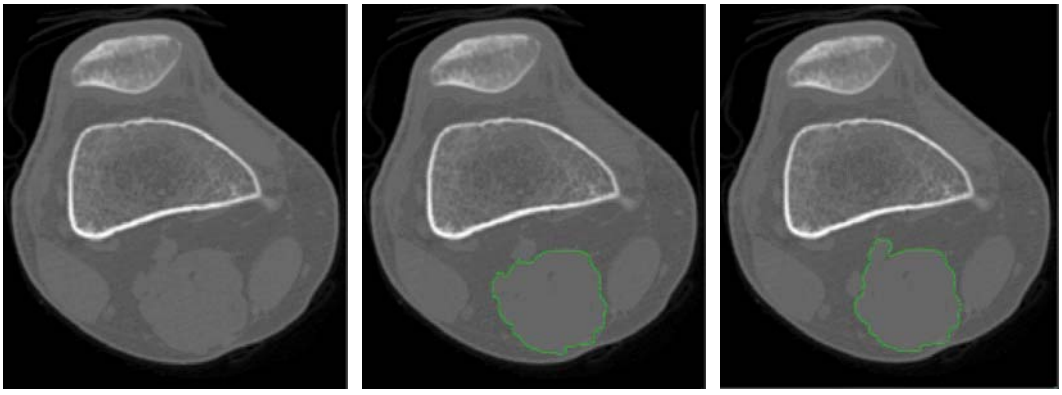
T_3: The same operator segments the same object once in two corresponding scenes in S_i and S_j $(i \neq j)$ by using method M (inter-scan).

For the given method of segmentation M, all possible pairs (O_1, 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* precision. A measure of precision for method M in a trial that produced segmented objects O_1 and O_2 for situation T_i is given by

$$PR_M^{T_i}(O_1, O_2) = \frac{\left| C_{O1} \bigcap C_{O2} \right|}{\left| C_{O1} \bigcup C_{O2} \right|}. \tag{5}$$

$PR_M^{T_i}(O_1, O_2)$ represents the total amount of the tissue that is common to both O_1 and O_2 as a fraction of the total amount of tissue in the union of O_1 and O_2. $PR_M^{T_i}(O_1, O_2)$ values estimated over the scenes in S_1, S_2,….,S_n utilizing operators H_1, H_2,….,H_m characterize the intra-operator, inter-operator, and repeat-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 determining how much the volumes (a commonly used figure of merit) of O_1 and O_2 agree will not be a robust measure of precision as illustrated in Figure 5. This is because O_1 and O_2 may have identical volumes but may constitute substantially different delineations.



(a) **(b)** **(c)**

FIGURE 5: Segmented objects (muscles) obtained in two different situations T_i and T_j ((b), (c)) in a slice of a CT scene of a knee ((a)). The two segmentations have nearly identical volumes, still they differ substantially.

2.5 Assessment of Accuracy

Let S_{td} be the set of scenes containing "true" delineations for the scenes in S. For any scene $C \in S$, let C_O^M be the scene representing the fuzzy object defined by an object O of B in C obtained by using method M, and let $C_{td} \in S_{td}$ be the corresponding scene of "true" delineation, all under the application domain <A, B, P>. The following measures are defined to characterize the accuracy of method M under <A, B, P> for delineation.

$$\text{False Negative Volume Fraction: } FNVF_M^d(O) = \frac{\left|C_{td} - C_O^M\right|}{C_{td}} \quad . \tag{6}$$

$$\text{False Positive Volume Fraction: } FPVF_M^d(O) = \frac{\left|C_O^d = C_{td}\right|}{\left|C_{td}\right|} \quad . \tag{7}$$

$$\text{True Positive Volume Fraction: } TPVF_M^d(O) = \frac{\left|C_O^M \bigcap C_{td}\right|}{\left|C_{td}\right|} \quad . \tag{8}$$

The meaning of these measures is illustrated in Figure 6 for the binary case. They are all expressed as a fraction of the volume of "true" delineation. $FNVF_M^d$ indicates the fraction of tissue defined in C_{td} that was missed by method M in delineation $FPVF_M^d$ denotes the amount of tissue falsely identified by method M as a fraction of the total amount of tissue in C_{td}. $TPVF_M^d$ describes the fraction of the total amount of tissue in C_{td} with which the fuzzy object C_O^M overlaps. Note that the three measures are independent; that is, none of them can be derived from the other two. True negative volume fraction has no meaning in this context since it would depend on the rectangular cuboidal region defining the scene domain C. Figure 7 presents an example showing the three factors for the application domain of brain parenchymal volume estimation via MRI T2 and PD scenes and by using the fuzzy connectedness method [13-15]. The surrogate of truth is obtained by manual delineation, and the estimates are based on binary objects.

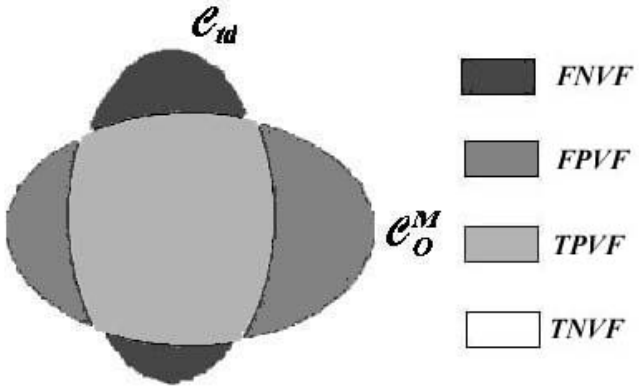
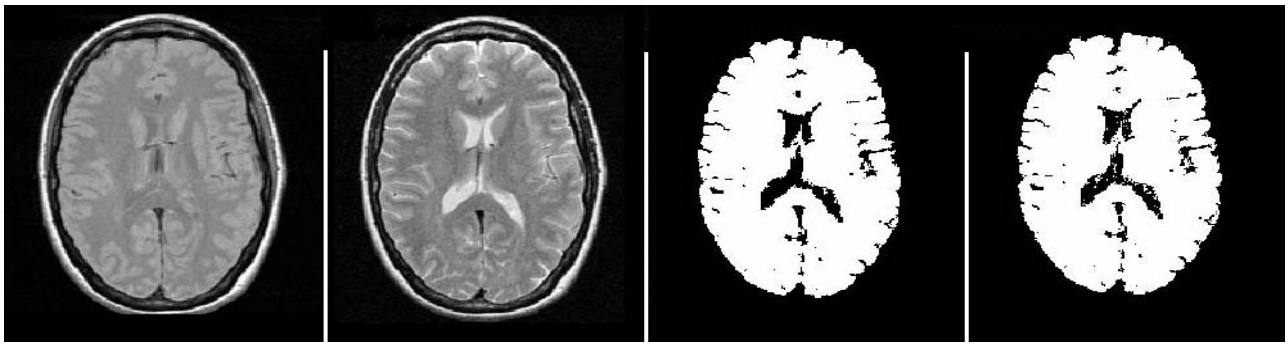


FIGURE 6: A geometric illustration of the three precision factors for delineation.



(a) **(b)** **(c)** **(d)**

FIGURE 7: Assessment of accuracy of segmentation. (a), (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 tracing. For this example, *FNVF* = 1.9%, *FPVF* = 0.2%, *TPVF* = 97.8%.

The following two measures are used to characterize the recognition accuracy of method *M* under <A, B, P>.

$$TPVF_M^r = \frac{\left| C_{tr} \bullet C_O^M \right|}{\left| C_{tr} \right|}, \tag{9}$$

$$FNVF_M^r = \frac{\left| C_{tr} \circ C_O^M \right|}{\left| C_{tr} \right|}. \tag{10}$$

It is very difficult for any segmentation method *M* to capture the weight information contained in C_{tr} associated with landmarks/features. Therefore, $FNVF_M^r$ cannot be assessed meaningfully. Further, it seems reasonable to determine what portion of the spread region of the landmarks/features is captured by *O*. The total weight in this captured region as a fraction of the total weight in C_{tr} defines $TPVF_M^r$ for characterizing the goodness of the qualitative (recognition) aspect of segmentation by method *M*. Analogously, $FNVF_M^r$ specifies the fraction of the total weight in C_{tr} that is missed by method *M*.

The accuracy of any two segmentation methods M_1 and M_2 for each of the five factors can be compared by comparing the set of values of the factor obtained for the scenes in *S* by using a paired *t*-test. For each factor, the 95% confidence interval (CI) may be computed to express the range of variation of this factor.

2.5 Assessment of Efficiency

Efficiency refers to the practical viability of a segmentation method. Two factors need to be considered to fully characterize efficiency: computational time and the human operator time required to complete segmentation of each study in a routine setting in the application domain <A, B, P>. We shall denote these two factors by E_M^c and E_M^h, respectively, for method *M*.

To assess E_M^c, the following two factors should be considered: (c_1) computational time required for algorithm training, (c_2) computational time for segmenting each scene. Most methods require an initial one-time "training" to

determine optimal parameter settings for the algorithm in the application domain <A, B, P>. The first component c_1 above corresponds to this aspect of computation. Let $t_M^{c_1}$ denote the total computational time needed for this one-time training. Then,

$$E_M^{c_1} = \frac{1}{t_M^{c_1} + L_1} \tag{11}$$

describes the first component of computational efficiency. Here L_1 is a constant. Its purpose is to handle methods, which do not require algorithmic training $\left(t_M^{c_1} = 0\right)$. L_1 should be chosen to be sufficiently small so that there is enough "resolution" in the resulting $E_M^{c_1}$ values to distinguish among methods. Alternatively, the methods can be compared directly based on $t_M^{c_1}$ values. The second component c_2 refers to the total computational time required (including any per-study training needed) to segment each scene. Let $t_M^{c_2}$ denote this time. Then,

$$E_M^{c_2} = \frac{1}{t_M^{c_2} + L_2} \tag{12}$$

describes the second component of computational efficiency, where L_2 is a constant. The purpose of L_2 is similar to that of L_1 (for completely manual segmentation, $t_M^{c_2} \approx 0$).

To assess E_M^h, the following three factors should be considered: (h_1) operator training time, (h_2) algorithm training, (h_3) and operator time per segmentation. In a clinical setting, for the routine use of a method, typically a technician runs the software that implements the method. Among methods, there is considerable variation in their complexity and intuition for use. The first component in h_1 above expresses the efficiency of method *M* from the point of view of ease of training of a medical technician for the routine clinical use of method *M*. Let $t_M^{h_1}$ denote the total time required to train each of a set of technicians for method *M*. then,

$$E_M^{h_1} = \frac{1}{t_M^{h_1} + K_2} \tag{13}$$

describes the first component of E_M^h. Here K_1 is a constant similar to L_1. In the second component h_2 of E_M^h, we consider the degree of operator help required for one-time initial training of the algorithm. Let $t_M^{h_2}$ denote the total operator time required for training for method *M* for each of a set of operators. Then

$$E_M^{h_2} = \frac{1}{t_M^{h_2} + K_2} \tag{14}$$

describes the second component of E_M^h. Here K_2 is a constant analogous to L_1. Most algorithms require per-study (for each scene under <A, B, P>) help from an operator for initialization (specifying initial boundary, seeds, initial segmentation, identifying landmarks) and/or for training to set values of the parameters of the method. The third component h_3 of E_M^h considers the extent of this help required in segmentation. This component is the most crucial

among all factors affecting efficiency since it represents the human effort needed to segment each scene. Let $t_M^{h_3}$ denote the total operator time required (including time for algorithm training) for segmenting each scene in S for method M. Then

$$E_M^{h_3} = \frac{1}{t_M^{h_3} + K_3} \tag{15}$$

describes the third component of E_M^h. The role of K_3 is analogous to that of L_1. This is for handling methods that do not require human intervention.

3. HOW TO COMPARE METHODS

The procedure for comparing two methods M_1 and M_2 under a given <A, B, P> consists of the following steps.

(1) Collect scenes $S_1 (= S), S_2, ..., S_n$ corresponding to n repeat scans of each of the scenes in S acquired for <A, B, P>. Produce scenes S_{td} representing surrogate of true delineations for the scenes in S.

(2) For methods M_1 and M_2, have operators $H_1, H_2, ..., H_m$ repeat segmentations of scenes in S. Have one operator segment scenes in $S_1 (= S), S_2, ..., S_n$ for methods M_1 and M_2.

(3) For i = 1, 2, 3, determine all possible values of $PR_{M_1}^{T_i}$ and $PR_{M_2}^{T_i}$.

(4) Knowing S_{td} and the segmentations of S produced by the operators, compute $FNVF_{M_j}^d$, $FPVF_{M_j}^d$, $TPVF_{M_j}^d$, $TPVF_{M_j}^r$, and $FNVF_{M_j}^r$, for $j = 1, 2$.

(5) Record $t_{M_j}^{c1}$, $t_{M_j}^{c2}$, $t_{M_j}^{h1}$, $t_{M_j}^{h2}$, $t_{M_j}^{h3}$ for $j = 1, 2$ during the segmentation experiments, and from these compute the respective efficiency parameters.

(6) For each method M_j, we get a set of values for each of the 13 parameters: $PR_{M_j}^{T_1}$, $PR_{M_j}^{T_2}$, $PR_{M_j}^{T_3}$, $FNVF_{M_j}^d$, $FPVF_{M_j}^d$, $TPVF_{M_j}^d$, $TPVF_{M_j}^r$, $FNVF_{M_j}^r$, $E_{M_j}^{c1}$, $E_{M_j}^{c2}$, $E_{M_j}^{h1}$, $E_{M_j}^{h2}$, $E_{M_j}^{h3}$. There are several choices for the statistical analysis of the 13 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 13 parameters for each method M_j by a weighted sum, the weight reflecting the importance given to that parameter for <A, B, P> and then do a paired t-test of the resulting single parameter.

(c) Do analysis of variance considering all 13 parameters.

4. CONCLUDING REMARKS

(1) The factors describing precision, accuracy, and efficiency are all essential in assessing the performance of segmentation methods. Most published methods of evaluation have ignored a majority of these factors.

(2) The question "Is method M_1 better than M_2 under <A, B, P>?" cannot be answered by a simple "yes" or "no". A descriptive answer in terms of the 13 parameters gives a more meaningful and complete assessment of the methods.

(3) Since binary results are produced by most segmentation methods in practice, and only delineation is considered, we suggest that, at a minimum, the following set of 8 parameters be evaluated: $PR_M^{T_1}$, $PR_M^{T_2}$, $PR_M^{T_3}$, $FPVF_M^d$, $FNVF_M^d$, $TPVF_M^d$, $E_M^{c_2}$, $E_M^{h_3}$ for any given method *M*.

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

(5) We are not aware of any attempt in the past to incorporate into the evaluation method the aspect of how well key features of an object that are considered important for <A, B, P> are captured in the segmentation. We are able to include this qualitative aspect of recognition also within the same common framework of evaluation.

(6) Most published methods have ignored the efficiency factor. The five components of efficiency are essential, E_M^{h3} 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* and/or *FPV*F and low *TPVF* 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 unless the method's efficiency is proven to be 100% (for all 5 factors) over a large (essentially infinite) number of data sets.

(7) The factors describing precision, accuracy, and efficiency are interdependent. To simultaneously improve all three factors for a method is difficult and requires considerable research. An attempt to increase accuracy may be accompanied by a decrease in efficiency and/or precision. These assertions are illustrated in Figure 8, wherein <A, B, P> is the application domain considered in Figure 7. Here thresholding using a fixed threshold value is the segmentation method *M*. Obviously $PR_M^{T_1}$ and $PR_M^{T_2}$ are both 100%. However, with repeat scan (Figures 8 (a), (b)) there is much variation in the result (Figures 8 (c), (d)) and $PR_M^{T_3}$ = 70.2%. The "true" delineations for the two scans of Figures 8 (a) and (b) are shown in Figures 8(e) and (f), respectively. It is clear that, although this method has high precision (except for the third factor $PR_M^{T_3}$) and efficiency, its accuracy is poor: $FNVF_M^d = 14.2\%$, $FPVF_M^d = 9.6\%$, and $TPVF_M^d = 76.1\%$. A possible way of improving accuracy is to modify *M* by having a human operator correct the results. This will of course bring down both efficiency and precision.

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

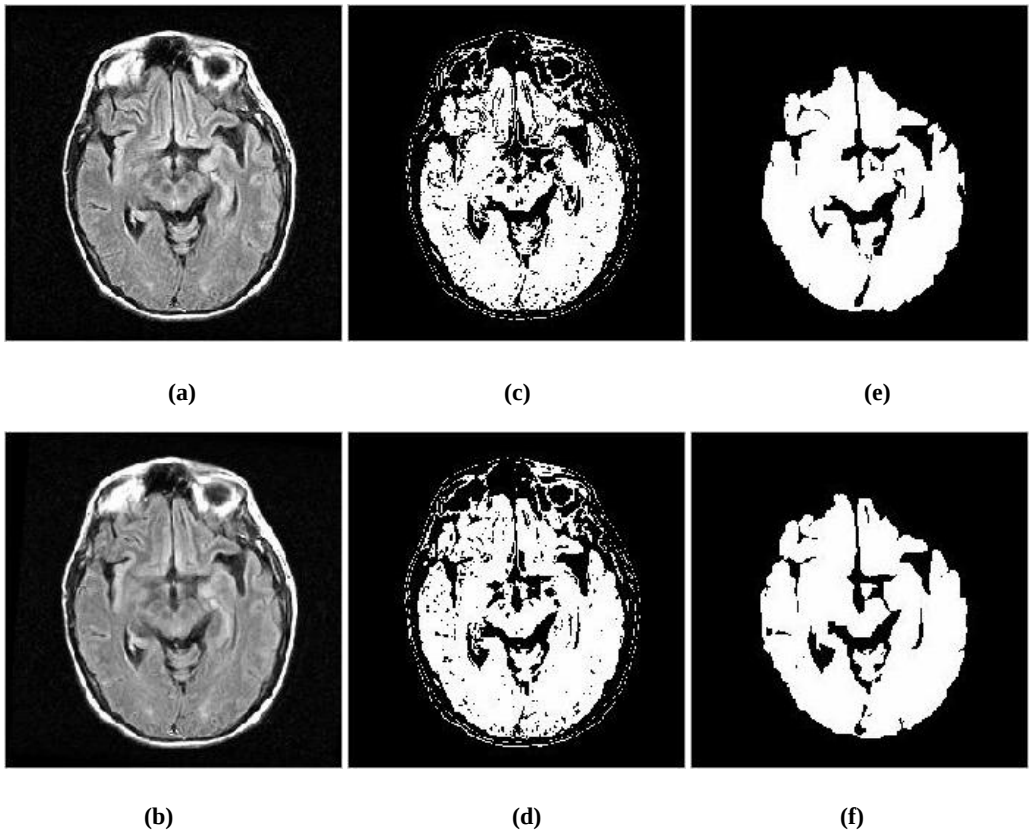


FIGURE 8: (a), (b). Two corresponding slices after registration of a pair of repeat scans (with a short time gap in between scans) of a patient's brain. (c), (d). Segmentation of (a) and (b) by fixed thresholding. The object of interest is brain parenchyma. (e), (f). "True" segmentations of (a) and (b).

ACKNOWLEDGEMENT

This work is supported by DHHS grants LM O-3502-MOD3, NS37172, and AR46902.

REFERENCES

[1] S.M. Pizer, E.P. Amburn, J.D. Cromaire, R.A. Geselowitz, T. Green, B.H. Romeny, J.B. Zimmerman and K. Zuiderveld: "Adaptive Histogram Equalization and its Variations," *Computer Vision Graphics and Image Processing,* 39:355-368, 1987.

[2] B. Morse, S. Pizer and D. Fritsch: "Robust Object Representation Through Object-Relevant Use of Scale," *SPIE Proceedings*, 2167:143-150, 1992.

[3] G. Gerig, O. Kubler, R. Kikinis and F. Jolesz: "Nonlinear Anisotropic Filtering of MRI Data," *IEEE Transactions on Medical Imaging*, 11:221-233, 1992.

[4] C. Pelizzari, G. Chen, D. Spelbring, R. Weichselbaum and C. Chen: "Accurate Three-Dimensional Registration of CT, PET and MR Images of the Brain," *Journal of Computer Assisted Tomography*, 13:20-26, 1989.

[5] N.R. Pal and S.K. Pal: "A Review of Image Segmentation Technique," *Pattern Recognition*, 26:1277-1294, 1993.

[6] Y. Chalana and Y. Kim: "A Methodology for Evaluation of Boundary Detection Algorithms on Medical Images," *IEEE Transactions on Medical Imaging*, 16:642-652, 1997.

[7] F. Mao, J. Gill and A. Fenster: "Technique for Evaluation of Semi-Automatic Segmentation Methods," *SPIE Proceedings*, 3661:1027-1036, 1999.
[8] G.T. Herman, S. Srihari and J. Udupa: "Detection of Changing Boundaries in Two- and Three-Dimensions," *Proceedings of the Workshop on Time Varying Imagery*, (eds.) N.I. Badler, J.K. Aggarwal, University of Pennsylvania, Philadelphia, Pennsylvania, pp. 14-16, April 1979.
[9] J.K. Udupa, S.N. Srihari and G.T. Herman: "Boundary Detection in Multidimensions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4:41-50, 1982.
[10] J.K. Udupa: "Interactive Segmentation and Boundary Surface Formation for 3-D Digital Images," *Computer Graphics and Image Processing*, 18:213-235, 1982.
[11] A. Falcao, J.K. Udupa, S. Samarasekera, S. Sharma, B.E. Hirsch and R. Lotufo: "User-Steered Image Segmentation Paradigms: Live-Wire and Live-Lane," *Graphical Models and Image Processing*, 60(4):233-260, 1998.
[12] A.X. Falcao and J.K. Udupa: "A 3D Generalization of User-Steered Live Wire Segmentation," *Medical Image Analysis*, 4:389-402, 2000.
[13] J.K. Udupa and S. Samarasekera: "Fuzzy Connectedness and Object Definition: Theory, Algorithms, and Applications in Image Segmentation, " *Graphical Models and Image Processing*, 58(3):246-261, 1996.
[14] P.K. Saha and J.K. Udupa: "Scale-Based Fuzzy Connected Image Segmentation: Theory, Algorithm and Validation," *Computer Vision and Image Understanding*, 77(2):145-174, 2000.
[15] P.K. Saha and J.K. Udupa: "Relative Fuzzy Connectedness Among Multiple Objects: Theory, Algorithms, and Applications in Image Segmentation," *Computer Vision and Image Understanding*, 82(1):42-56, 2001.
[16] J.K. Udupa, L. Wei, S. Samarasekera, Y. Miki, M.A. van Buchem and R.I. Grossman: "Multiple Sclerosis Lesion Quantification Using Fuzzy-Connectedness Principles," *IEEE Transactions on Medical Imaging*, 16(5):598-609, 1997.
[17] J.K. Udupa, B.E. Hirsch, S. Samarasekera, H. Hillstrom, G. Bauer and B. Kneeland: "Analysis of *In Vivo* 3D Internal Kinematics of the Joints of the Foot," *IEEE Transactions on Biomedical Engineering*, 45911):1387-1396, 1998.
[18] P.K. Saha, J.K. Udupa, E. Conant, D.P. Chakraborty and D. Sullivan: "Breast Tissue Glandularity Quantification via Digitized Mammograms," *IEEE Transactions on Medical Imaging*, 20(8):792-803, 2001.
[19] T. Lei, J.K. Udupa, P.K. Saha and D. Odhner: "Artery-Vein Separation via MRA – An Image Processing Approach," *IEEE Transactions on Medical Imaging*, 20(8):689-703, 2001.
[20] P.K. Saha, J.K. Udupa and J.M. Abrahams: "Automatic Bone-Free Rendering of Cerebral Aneurysms via 3D-CTA," *SPIE Proceedings*, 4322:1264-1272, 2001.
[21] J.K. Udupa, D. Odhner and H.C. Eisenberg: "new Automatic Mode of Visualizing the Colon via Cine CT," *SPIE Proceedings*, 4319:237-243, 2001.
[22] J.G. Liu, J.K. Udupa, D. Hackney and G. Moonis: "Brain Tumor Segmentation in MRI Using Fuzzy Connectedness Method," *SPIE Proceedings*, 4322:1455-1465, 2001.