

医学图像分割算法的评价方法

张石 董建威 余黎煌

(东北大学信息科学与工程学院, 沈阳 110004)

摘要 对医学图像分割算法的客观评价是推进算法在临床上得到应用的关键。针对目前对医学图像分割方法的研究较多,而对分割算法的评价方法的研究却很少的问题,提出了一种判断和比较医学图像分割算法优劣的评价方法。首先对现有的几种评价方法进行了综述,并总结出了一套评价系统。可靠性、精确性、区域统计特性和效率是评价一个分割方法的4个重要参数,结合医学图像分割分别对它们的定义进行了说明。这些参数互相影响,评价分割算法时必须权衡这些指标,根据不同的应用背景赋予它们不同的权重。此外,还介绍了如何建立医学图像分割金标准数据库的方法。最后,通过 Insight Toolkit(ITK)软件包中的两个算法例子,结合脑白质分割的医学背景,演示了如何利用本文评价系统来对这两种分割算法做出比较。实验结果表明,该评价方法可行,比较结果具有合理性。该研究为医学图像分割算法的评价提供了科学合理的方法,同时也指出了推动医学图像分割算法在临床上应用所应解决的问题。

关键词 图像分割 分割评价 医学图像 ITK 金标准

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2009)09-1872-09

The Methodology of Evaluating Segmentation Algorithms on Medical Image

ZHANG Shi, DONG Jian-wei, SHE Li-huang

(School of Information Science and Engineering, Northeastern University, Shenyang, 110004)

Abstract Objective evaluation of medical image segmentation algorithms is one of the important steps toward establishing validity and clinical applicability of an algorithm. Since there are a large number of articles presenting segmentation methods on medical image, with few studying the evaluation methods on their performance, this paper presents an evaluation method for different segmentation algorithms. The author first gives a survey of several available evaluation methods and presents a systematic summary. Reliability, precision, region statistical characteristics and efficiency are the four most important metrics. The definitions of them are then described based on the image segmentation process. For comparison, weights should be added to these metrics according to the application. Moreover, the author also presents a method on how to construct gold standard of medical images. At last, with the task of brain white matter segmentation, the author demonstrates how to make use of the proposed evaluation method to compare two segmentation algorithms in insight toolkit(ITK). The experiment results show that this method is practical and reasonable. This study gives a scientific method for the evaluation of segmentation algorithms on medical image. Meanwhile, it points out the problems to be solved before the segmentation algorithms could be put into use in clinic.

Keywords image segmentation, segmentation evaluation, medical image, insight toolkit(ITK), gold standard

1 引言

图像分割是指将图像中感兴趣区域(ROI)的边界找到,使得边界内部和外部的像素分别具备相似的特征(强度、纹理等)。医学图像分割是对医学图像进行其他后续处理的基础。对图像中目标区域的准确分割对于计算机辅助诊断、制订手术计划、目标3维重建以及放射性治疗评价等都具有重要意义。近几十年来,随着医学成像设备的不断完善,医学图像的分割算法也层出不穷,但是很少能够在临床上得到广泛应用。用一个全面的医学图像数据集集合来客观评价医学图像分割算法是将该算法向临床应用推进的关键一步^[1]。然而,目前对某一医学图像分割算法的评价,大部分情况都是通过目测分割结果的方法来给出定性的评价,很难给出客观的、系统的、定量的评价。究其原因,主要存在如下两个问题:

(1)不同组织或器官,不同成像模式(CT/MRI/PET/US)医学图像的特点差别很大,一种医学图像分割方法很可能仅仅对一种图像分割有效,甚至具体到一个特定的部位,无法得到一个通用的分割方法,得到算法的客观评价则更难。

(2)对于一幅特定的医学图像,不同的医学图像专家很可能给出不同的分割结果,存在因人而异的特点,因此,很难对计算机分割方法得到的分割结果做一客观评价,缺少关于医学图像分割的金标准数据库。

图像分割算法的评价方法受到关注要远远落后于对分割算法本身的研究,1977年,Yasnoff等人率先提出了“错误分类百分比”和“像素距离误差”两个测度,来评价图像分割算法^[2]。1986年,Canny从图像的边缘检测角度,提出了3个最优边缘检测准则:(1)不漏检真实存在的边缘,也不把非边缘点作为边缘点检出;(2)检测到的边缘点与实际边缘的距离最小;(3)实际存在的边缘点与检测到的边缘点一一对应^[3]。1996年,章毓晋对前人的大量工作进行了总结,将评价方法归结为两类——直接法(分析法)和间接法(实验法),并对每种方法进行了简要介绍,但其文中却没有提及在研究一个特定的分割任务时,尤其是医学图像分割任务时,如何去选择和合适的评价方法^[4-5]。1997年,Chalana

等人提出了一种通过比较计算机分割结果与多个专家手工标记结果,来评价医学图像分割算法的方法,并用两个超声图像分割的例子演示了如何应用其提出的评价方法^[1]。2006年,Anon提出了一种利用支持向量机(SVM)学习方法对分割算法进行无监督评价的方法^[6]。同年,Udupa等人提出了一个针对医学图像分割方法评价的框架,他将图像分割问题划分为目标识别和描述,并提出3个要素——可靠性、精确性和效率,列举了很多医学图像分割的例子,并介绍了各个评价参数的具体计算方法^[7]。同年,Ortiz等人提出了一种新的由3个参数组成的测度方法,这3个参数是:正确分类的比例、过度分割的比例和欠分割的比例^[8]。国内在这方面的研究较少,且没有专门围绕医学图像分割展开讨论的文献。2000年,侯格贤等人将模糊集合的概念应用到分割评价方法中,引入模糊度测度用于评价图像分割的质量^[9]。其他研究者也在该领域开展了一些研究工作,但都并不深入,专门研究医学图像分割算法的评价方法的文献则几乎没有。在对上述文献进行分析综合后,得到如图1所示的医学图像分割评价系统的总体运行流程。按照图1所示的各个环节,对国内外目前研究的几种评价方法的基本思想做一介绍,重点研究在面临一个特定的分割任务(例如人脑MRI图像中脑白质的分割)时,如何构建一个分割评价系统以客观评价和比较所采用的分割算法。

2 分析法与实验法

分析法是通过研究分割算法本身的原理和性质来对分割算法进行评价。评价指标包括:对先验知识的利用程度(如利用解剖图谱的分割),算法的运行方式(串行、并行),算法的空间复杂性与时间复杂性以及算法的抗噪性和稳定性等。分析法评价的意义在于发现算法的实质性缺陷,并明确算法改进的方向,特别适合于定性评价。

实验法则是通过实现一个分割算法的实例,通过对分割的结果进行研究,来评价该分割算法。它又分为有监督评价法和无监督评价法。有监督评价法是建立在与可供参考的分割标准进行比较的基础上,获得一系列的误差数值。而无监督法则借助分割结果自身的统计特性来评价分割算法。

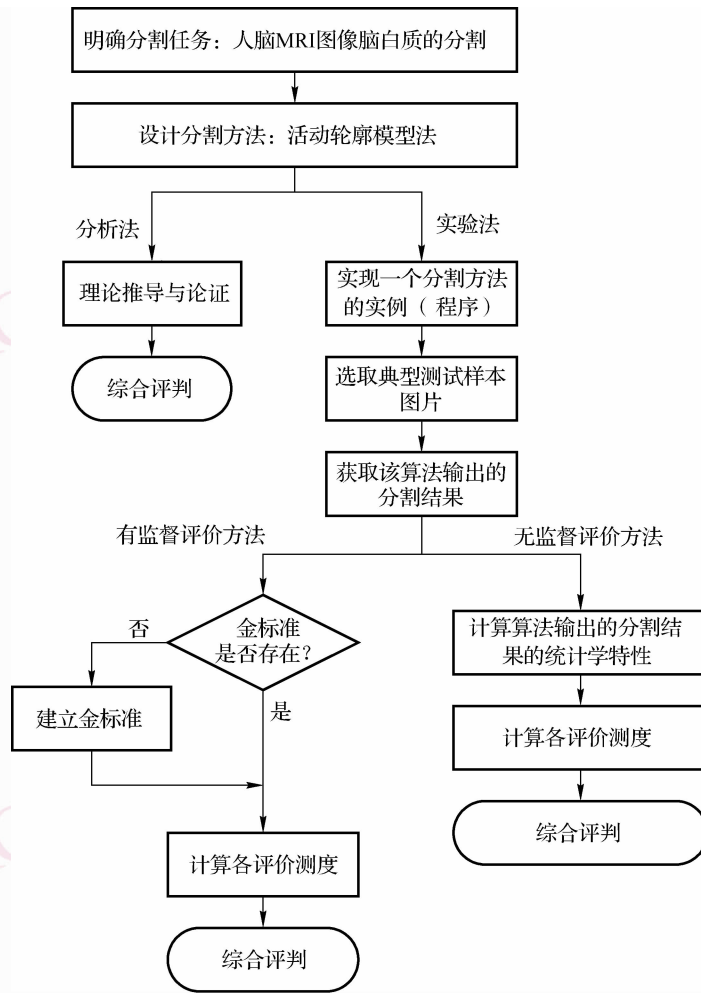


图 1 医学图像分割评价系统的运行流程

Fig. 1 Process for evaluating segmentation algorithm on medical image

3 评价测度

评价测度就是衡量算法性能优劣的各种数学指标,目前能够查阅到的评价测度有很多,这里仅列举最典型的 4 个评价测度:可靠性、精确性、区域统计特性和效率。

3.1 可靠性

可靠性是指利用统计学规律考虑图像分割的各个变化因素对图像进行重复分割,两次分割结果之间的接近程度;在评价分割算法的可靠性时,有如下几个变化因素需要考虑:(1)在成像过程中患者身体位置的不同,会导致图像(目标)的旋转等;(2)人机交互过程中手工操作的变化,如初始化位置的给定,会导致算法初始化状态的不同。可靠性可以通过如下 3 个方法来考察:

- (1) 同一个操作者对一幅图像中的特定目标使用相同的分割算法分割 n 次;
- (2) n 个操作者对同一幅图像中的特定目标使用相同的分割算法各分割一次;
- (3) 同一个操作者对分 n 次扫描的同一个部位的 n 幅图像中的特定目标使用同一种分割算法分别进行分割。

采用上述 3 种方法可获得 3 组分割结果,设 S_1, S_2, \dots, S_n 为其中一组分割结果(区域内像素集合),则采用如下可靠性计算公式:

$$P = \frac{|S_1 \cap S_2 \cap \dots \cap S_n|}{|S_1 \cup S_2 \cup \dots \cup S_n|} \quad (1)$$

其中,分子表示 n 次分割结果重叠像素的个数,分母表示 n 次分割结果得到的像素并集中元素的个数。得到的 3 个可靠性测度 P_1, P_2, P_3 分别代表了操作

者自身误差,操作者间误差和成像设备误差(严格来说 P_3 也包含操作者自身误差)。 P 值越接近于 1,代表算法对外界条件的适应性越强,可重复性就越好。

3.2 精确性

精确性是指算法分割结果与真实分割结果(金标准)的一致性程度。它有 4 种描述方法:像素法、区域法、点对法和边界法。像素法就是计算错误分类的像素数占参考像素总数的比重;区域法计算两个对应区域的重叠比例;点对法是在两个分割结果的边界上首先找到对应的点对,然后测量它们的一致程度;边界法则比较分割的边界与参考边界的贴近程度。分割问题可看作将一组像素点进行分类的问题,因此它可以被看成是一个聚类问题,用于测度一致性或相似性的聚类分析的方法也都可以用来衡量分割结果和金标准的贴近程度。

首先,可以用基于距离的测度来评价基于边界的分割方法。设分割算法得到的边界点的集合为 $B = \{b_i; i = 1, \dots, K\}$,金标准的边界点的集合为 $T = \{t_i; i = 1, \dots, N\}$,点到边界的距离定义为:

$$d(b_i, T) = \min_{n \in \{1, \dots, N\}} |b_i - t_n|$$

对于一幅图像,可以计算出如下 3 个基于距离的测度:

$$\text{平均绝对偏差: } MAD = \frac{1}{K} \sum_{i=1}^K d(b_i, T) \quad (2)$$

$$\text{最大绝对偏差: } MaxD = \max_{i \in \{1, \dots, K\}} \{d(b_i, T)\} \quad (3)$$

误差范围在 e 个像素以内的像素比例:

$$PE = \frac{|\{b_i \in B; d(b_i, T) < e\}|}{K} \quad (4)$$

此外,还可以采用基于区域(面积或体积)的测度。为简单起见,假设一幅图像仅包含一个目标区域,其他为背景,设集合 S 代表分割算法得到的目标区域的像素点集合,集合 T 代表金标准的目标区域的像素点集合, I 表示整幅图像的像素集合,则定义如下 4 个测度:

$$\text{真阳性: } TP = S \cap T \quad (5)$$

$$\text{假阳性: } FP = S - T \quad (6)$$

$$\text{假阴性: } FN = T - S \quad (7)$$

$$\text{真阴性: } TN = I - T - S \quad (8)$$

这 4 个测度可以用类似文氏图的形式表达,如图 2 所示。

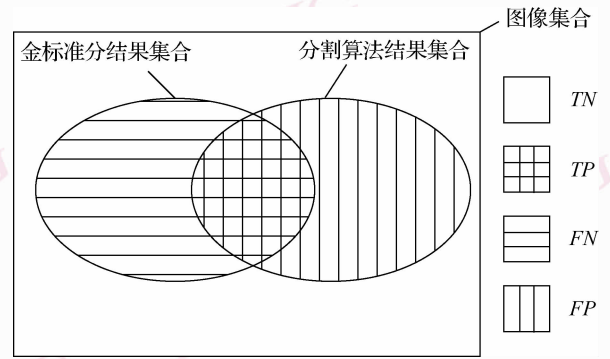


图 2 基于区域的精确性评价测度示意图

Fig. 2 Illustration of evaluation metrics for region-based algorithm

在此基础上再定义如下 4 个比例:

$$TPF = \frac{|TP|}{|T|} \quad (9)$$

$$FNF = \frac{|FN|}{|T|} \quad (10)$$

$$FPF = \frac{|FP|}{|I - T|} \quad (11)$$

$$TNF = \frac{|TN|}{|I - T|} \quad (12)$$

TPF 表示真阳性目标像素数占目标区域所有正确像素数的比例; FNF 表示漏分割目标像素数占目标区域所有正确像素数的比例; FPF 表示将背景像素错误地分割为目标像素的数目占背景正确像素数的比例; TNF 表示分割出来的背景像素数占正确背景像素数的比例;若图像区域的尺寸与目标区域的尺寸比例合适,则分割算法可以通过绘制 TPF 和 FPF 的接收者操作特性(ROC)曲线得到评价和优化^[10]。而错误比例通过 ROC 方法来评价不合适,可以通过下式评价^[11]:

$$FF = 1 - \frac{|FP| + |FN|}{|T|} \quad (13)$$

将上述结果中的普通集合换成模糊集合,得到的测度就是改进了的模糊集精确性测度^[7]。

3.3 区域统计特性

在有些情况下,分割结果没有金标准可供参考,因此产生了无监督评价方法。它采用区域的统计学特性(灰度均匀性、纹理特征等)作为测度。基本的测度包括如下 3 个:区域内均匀性、区域内差异性和区域间差异性。

区域内均匀性 (*uniformity of intra region*): 通过区域内像素特征 (如灰度) 的相似度计算区域的均匀性^[12-13], 公式为

$$UR = 1 - \frac{1}{|I|} \sum_{k=1}^N \frac{\sum_{s \in R_k} \left(g_I(s) - \frac{1}{|R_k|} \sum_{t \in R_k} g_I(t) \right)^2}{\left(\max_{s \in R_k} (g_I(s)) - \min_{s \in R_k} (g_I(s)) \right)^2} \quad (14)$$

这里: 设图像 I 被分割成 N 个区域 (集合), R_k 表示第 k 个区域 (集合), $|I|$ 代表图像 I 中像素的总数, $g_I(s)$ 对应像素 s 的灰度级强度函数或其他像素特征函数 (颜色、纹理等), $|R_k|$ 代表区域 R_k 中像素的总数。

区域内差异性: 通过计算区域内像素灰度的二阶导数的平方均值作为区域内差异性的测度^[12], 其式为

$$DR = \frac{1}{N} \sum_{k=1}^N \left(\frac{1}{|R_k|} \sum_{s \in R_k} \left(\frac{d^2 g_I(s)}{ds^2} \right)^2 \right) \quad (15)$$

但是, 该测度对于以纹理为区域特征的图像分割并不适用, 对于纹理分割的区域内差异性测度, 可参见文献[14]。

区域间差异性: 通过计算两区域间的不一致性, 获得区域间差异性^[12], 其式为

$$DIR = \frac{1}{C_N^2} \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^N |f_i(R_i) - f_j(R_j)|}{\max_{s \in I} (g_I(s)) - \min_{s \in I} (g_I(s))} \quad (16)$$

其中, C_N^2 为区域组合数, $g_I(s)$ 为像素灰度特征函数, $f_i(R_i)$ 为区域特征函数, 一般为区域平均灰度。其他无监督评价测度可参见文献[1]和文献[4]。

3.4 效率

效率是指计算机和用户参与的用于算法训练和执行的总时间。完成一次分割任务所用时间也是一项重要的测度, 可以分为人工操作时间和计算机运算时间两部分。人工操作时间又分为初始化算法时间和训练算法时间。计算机运算时间又分为算法的学习 (训练) 时间和算法运算时间。

值得一提的是, 效率不仅仅和采用的算法本身有关, 也和采用的算法的实现方法 (如 Matlab, C++ 等) 有关, 与采用的软硬件平台 (操作系统、处理器、内存容量等) 有关。因此, 在描述算法效率的同

时, 也必须明确算法的实现方法及软硬件的资源情况。

4 综合评价系统

实际上, 前面几节介绍的这些测度之间并不是互相独立的, 它们互相影响, 而且通常很难在保证其他几个参数不变坏的情况下改善其中一个参数。对分割算法的评价必须权衡这些指标, 根据不同的应用背景赋予它们不同的权重。

综上所述, 一个完整的分割评价系统包含如下 5 个方面的内容: (1) 一套容易计算的、高效的、有意义的测度。(2) 若干幅具有代表性的、真实的临床采集图像。(3) 可以用来作为分割参考的对应的金标准。(4) 一些用于测试和比较的具有标准输入输出的分割算法模块。(5) 一个可以容纳分割算法和评价方法的软硬件系统。在建立这样一个评价系统前, 还必须明确其应用范围。应用范围包括如下 3 个方面的内容: (1) 一个任务: 例如肿瘤的分割; (2) 分割的部位: 例如大脑; (3) 图像类型: 如 MRI 2 维图像。只有满足了上述这些条件, 对医学图像分割算法的评价才是客观的和有意义的。

5 金标准数据库的建立

严格来说, 在很多情况下, 绝对客观的金标准是不可能获得的, 也是不可知的。为此, 通常采用如下 3 种方法来获取用做分割参考的金标准。

(1) 人工标记法 感兴趣对象的边缘或区域由几位有经验的专家手工标记。为避免专家们之间分割的不一致性, 可采用平均法来获得最终的金标准, 文献[15]提出了一种期望最大化算法, 也可用来解决不一致性的问题。人工标记法得到的分割结果往往都是一个二值的分割结果, 即某个点或者是目标区域中的点, 或者不是目标区域中的点。对于边界模糊的图像, 无法给出像模糊集中的隶属度那样的一个隶属度函数 (0~1 之间的一个小数)。人工标记法还具有如下几个缺点: (1) 需要耗费专家很多的时间和精力, 成本较高; (2) 易变性, 不同的专家或同一个专家在不同时间对同一幅图像所得到的多个分割结果往往不一致, 具有可变性; (3) 专家在观看图像时, 图像边缘的模糊性, 显示设备的设置以及

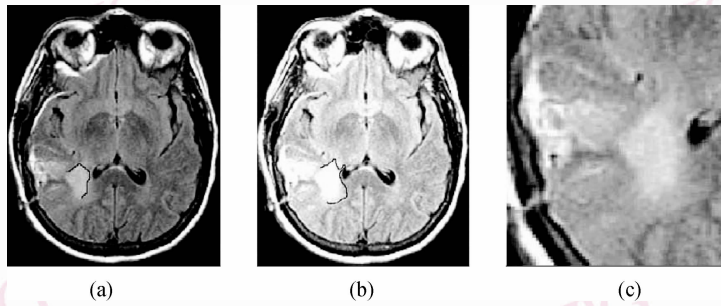


图3 一个手工分割实例

Fig. 3 An example of manual segmentation

专家的视觉特征都会影响手工分割的结果。如图3所示为一个此类图像示例^[7]。图3(a)与图3(b)具有不同的屏幕对比度,它们与图3(c)具有不同的放大倍数,导致目标边界标记不一致。

(2)数学建模法 首先设计一个二值化的仿真图像,其中明确定义了目标区域和背景。然后,加入组织间的局部体效应影响和成像过程中混入的噪声等,最终生成一幅仿真的医学图像。这种方法的缺点显而易见,采用的图像不是通过影像设备获取的临床真实图像,必然与其存在差异,说服力不强。

(3)成像过程仿真法 制作与人体组织或器官的密度和结构相似的人造模型,然后将其放入成像设备进行成像,则得到的图像即为已知金标准的图像。人造模型也可以用第2种方法的数学模型代替,成像设备的成像过程可以通过仿真方法生成投影数据,然后再进行反投影变换和图像重建。著名的BrainWeb的MRI simulator采用的就是该方法,它允许用户独立地指定成像参数来生成期望的大脑MRI图像^[16-17]。这种方法的缺点在于人造模型的成本很高,且无法完全符合人体的组织结构,采用的数学模型获得的图像与真实图像的差别也较大。

从对上述3种方法的分析可见,采用人工标记法的金标准更符合医学图像分割这一任务的终极目标。将所有的外部可变因素标准化(显示设备的亮度、对比度,图像的放大倍数等)^[18-19],将目标区域边界的过渡区以模糊概率的形式而不是0-1的形式表达,这样建立的金标准将具有更大的参考价值。

为了推动医学图像分割方法的研究,一些组织和机构为了解决研究者难以获得金标准数据库的问题,

专门建立了针对不同分割任务的金标准数据库,其中以国际脑分割数据库(IBSR)^[20]和BrainWeb^[17]最有影响力,如图4所示为其典型样本及金标准,图4(a)来自IBSR,图4(b)来自BrainWeb,具体内容请参见文献[17]、[20]相关介绍。

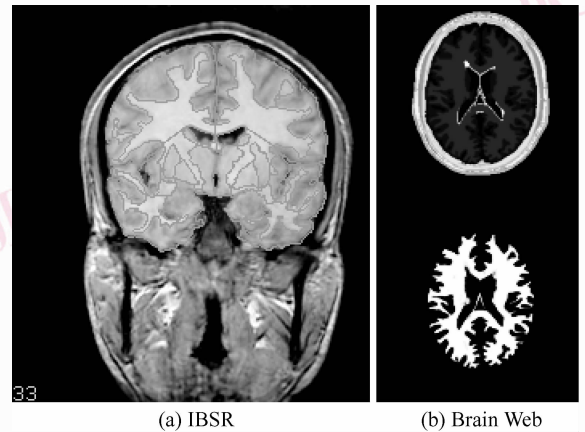


图4 脑MRI图像及脑白质分割金标准

Fig. 4 Brain MRI image and its gold standard of white matter

医学影像参考标准数据库的建立也是目前医学图像处理研究急需解决的问题,在欧盟医学信息学-医学图像处理工作组(EFMI WG MIP)2006年召开的“医学图像处理系统有效性”会议上,专门对参考图像数据库(RID)的现状、未来和研究计划展开了讨论,并进一步明确了如何建立全面的RID以推动医学图像处理与分析的发展^[21]。

6 分割方法比较及评价实例

图像分割评价的目的主要有两个:一个是研究算法在不同分割情况中的表现,掌握如何选择和修正其参数以适应特定的分割任务;另一个是分析比

较多个分割算法在面对同一分割任务时的优劣程度以选取合适的算法^[5]。算法优化在前面提过可以用 ROC 曲线的方法。这里以脑 MRI 图像白质分割用于解剖结构研究为例,来评价两种不同的分割算法在完成这同一分割任务时的性能差异。采用的两种算法及其实现方法均为 NIH(national institutes of health)的用于研究医学图像分割和配准的 insight toolkit(ITK)软件开发包中的算法^[22]。采用的硬件平台为 AMD sempron 1.9 GHz CPU,1GB 内存。软

件平台为 Windows XP 操作系统,C++ 编程语言和 Visual C++6.0 编译器。

实验采用的测试图片为 IBSR 数据集中的男性 T1 加权脑 MRI 图像,编号为 788_6 的第 52 层。分割任务为脑白质区域的提取。两位操作者采用了 ITK 中的两个分割算法,共获得 6 个分割结果,具体实验方法及算法参数的选择如表 1 所示。原图及其金标准分割结果以及 6 个计算机获得的分割结果如图 5 所示。

表 1 实验方法及算法参数选择

Tab. 1 Experiment method and parameters corresponding to the algorithms

ShapeDetectionLevelSetFilter 算法			ConnectedThresholdImageFilter 算法			
实验序号	种子点坐标 (x,y)	梯度参数 δ	实验序号	种子点坐标 (x,y)	灰度范围	
操作者 1	1	(103,148)	2.6	4	(103,148)	[140 190]
操作者 2	2	(126,140)	2.6	5	(126,140)	[137 184]
操作者 3	3	(126,140)	2.4	6	(126,140)	[121 229]

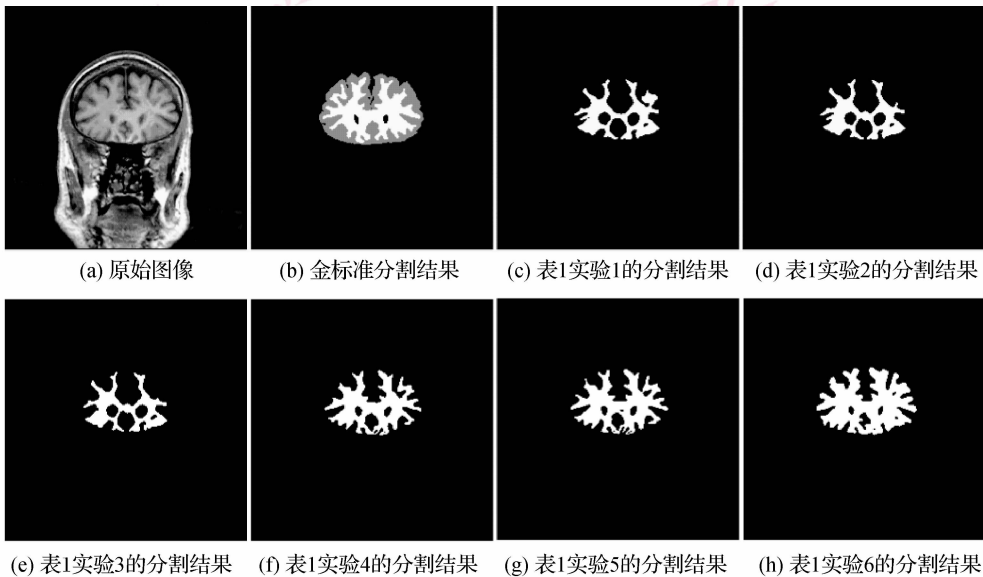


图 5 MRI 图像脑白质分割算法评价实例图片

Fig. 5 Segmentation results of white matter from Brain MRI

在可靠性方面,因对同一部位两次成像的数据无法获得,因此只测量了 P_1 和 P_2 。采用的这两种分割算法均属区域增长法,因此均需要人工指定种子点的位置和终止条件,因操作者自身和之间的差异性,会造成初始化条件等有所不同。实验结果如表 2 所示,从表中可以看出,ShapeDetectionLevelSetFilter 算法(以下简称为 S 算法)的可靠性高,但分割结果不精

确,且耗时较长。ConnectedThresholdImageFilter 算法(以下简称为 C 算法)可靠性稍差,但分割结果较前一算法精确,且计算时间短。

由上面的分析可知,两种算法各有其优缺点,若要对这两种方法做出最终比较,必须根据算法特点和实际应用背景建立综合评判模型。根据本分割任务,医生首先需要对算法的可靠性、精确性和效率设

表2 分割算法的测度数值
Tab.2 Metrics results of the segmentation algorithms

编号	P_1 (%)	P_2 (%)	TPF (%)	FNF (%)	FPF (%)	TNF (%)	算法初始化所用时间(s)	算法完成计算所用时间(s)
1			24.31	75.69	2.39	97.61	3	8.234
2	91.09	93.19	24.58	75.42	2.21	97.79	3	8.578
3			23.40	76.60	2.04	97.96	5	8.453
4			28.51	71.49	3.40	96.60	2	0.656
5	93.75	75.21	29.05	70.95	3.62	96.38	2	0.593
6			36.69	63.31	4.90	95.10	5	0.516

置一个底线,在此前提下按“效率 > 精确性 > 可靠性”的优先级排序,为了归一化各项权重,假设5s的算法运行时间等同于一个精度百分点或一个可靠性百分点,则得到下式:

$$C_{sc} = \mu \frac{\sum_{i=1}^3 (P_{Si} - P_{Ci})}{3} - \lambda \frac{(t_s - t_c)}{5} + \gamma \cdot \frac{(TPF_s + TNF_s) - (TPF_c + TNF_c)}{2} \quad (17)$$

取 $\lambda = 0.5$, $\gamma = 0.3$, $\mu = 0.2$ 并假设第2和第6组为最优化后的分割结果, P_3 等于 P_1 , 则计算得 $C_{sc} = -2.5882$, 说明S算法总体来说没有C算法好,这与前面的定性分析的结果是一致的,说明该综合评价公式具有一定的合理性,将测试样本的数量扩大,最终计算得的平均比较结果将更加具有客观性。

7 结 论

针对医学图像分割这一研究课题,提出了一套系统的对分割算法进行比较评价的方法,并结合ITK算法包,对该评价方法进行了实验,并对实验结果进行了简要分析。目前,绝大部分研究医学图像分割的文章,在实验分析部分均只给出分割前后的图像,只能用肉眼评价其分割结果,且不具有可比性。本文提出的评价系统,再配合大量的临床图像样本,对分割算法的比较以及推动医学图像分割算法在临床上的应用,将具有重大意义。此外,对研究其他图像分割及跟踪算法的评价方法也有一定的指导意义。

然而,面对医学图像分割问题的复杂性,加之算法实现方法的各异性,算法之间的比较往往比较困

难,因此,对评价系统的每一个环节进行标准化也是十分有必要的。需要标准化的内容主要包括:

- (1) 建立每种分割任务的参考数据集(如IBSR);
- (2) 规定统一的算法实现方法和实验平台(如ITK);
- (3) 设计统一的测度指标及统计分析软件包;
- (4) 明确在每种分割任务中各个测度的权重。

只有这样,才能真正推动医学图像分割算法的研究向前发展,使其在临床上得到应用。

参考文献 (References)

- 1 Chalana V, Kim Y. A methodology for evaluation of boundary detection algorithms on medical images [J]. IEEE Transactions on Medical Imaging, 1997, 16(5):642-652.
- 2 Yasnoff W A, Mui J K, Bacus J W. Error measures for scene segmentation [J]. Pattern Recognition, 1977, 9(4):217-231.
- 3 Canny J. A computational approach to edge detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 8(6): 679-698.
- 4 Zhang Y J. A survey on evaluation methods for image segmentation [J]. Pattern Recognition, 1996, 29(8): 1335-1346.
- 5 ZHANG Y J. A classification and comparison of evaluation techniques for image segmentation [J]. Journal of Image and Graphics, 1996, 1(2):151-158. [章毓晋. 图象分割评价技术分类和比较 [J]. 中国图象图形学报, 1996, 1(2):151-158.]
- 6 Anon. Adaptive evaluation of image segmentation results [A]. In: Proceedings of the 18th International Conference on Pattern Recognition [C], Hong Kong: Institute of Electrical and Electronics Engineers Inc, 2006: 399-402.
- 7 Udupa J K, LeBlanc V R, Ying Zhuge, et al. A framework for evaluating image segmentation algorithms [J]. Computerized Medical Imaging and Graphics, 2006, 30(2):75-87.
- 8 Ortiz A, Oliver G. On the use of the overlapping area matrix for image segmentation evaluation: A survey and new performance measures [J]. Pattern Recognition Letters, 2006, 27(16): 1916-1926.

- 9 Hou Ge-xian, Bi Du-yan, Wu Cheng-ke. Researches on evaluation methods for image segmentation [J]. Journal of Image and Graphics, 2000, **5**(1):39-43. [侯格贤, 毕笃彦, 吴成柯. 图象分割质量评价方法研究 [J]. 中国图象图形学报, 2000, **5**(1):39-43.]
- 10 Bommel J H, Musen M A. Medical Informatics [M]. Shanghai: Publisher of Shanghai Science and Technology, 2002: 227-230. [Bommel J H, Musen M A 著. 医学信息学[M]. 包含飞等译. 上海:上海科学技术出版社, 2002:227-230.]
- 11 Fenster A, Chiu B. Evaluation of segmentation algorithms for medical imaging [A]. In: Proceedings of the 27th Annual International Conference of the Engineering in Medicine and Biology Society[C], Shanghai: Institute of Electrical and Electronics Engineers Inc. 2005: 7186-7189.
- 12 Chabrier S, Emile B, Rosenberger C, *et al.* Unsupervised performance evaluation of image segmentation [J]. EURASIP Journal on Applied Signal Processing, 2006, **15**:1-12.
- 13 Levine M D, Nazif A M. Dynamic measurement of computer generated image segmentations [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1985, **7**(2): 155-164.
- 14 Chabrier S, Rosenberger C, Laurent H, *et al.* Evaluating the segmentation result of a gray-level image [A]. In: Proceedings of 12th European Signal Processing Conference[C], Vienna: European Association for Signal, Speech and Image Processing, 2004: 953-956.
- 15 Warfield S K, Zou K H, Wells W M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation [J]. IEEE Transactions on Medical Imaging, 2004, **23**(7): 903-921.
- 16 Kwan RK-S, Evans A C, Pike G B. MRI simulation-based evaluation of image-processing and classification methods [J]. IEEE Transactions on Medical Imaging, 1999, **18**(11): 1085-1097.
- 17 BrainWeb: Simulated Brain Database [EB/OL]. <http://www.bic.mni.mcgill.ca/brainweb/>, 2006-6-12/2007-6-17.
- 18 Pizer S M. Intensity mappings to linearize display devices [J]. Computer Graphics and Image Processing, 1981, **17**(3): 262-268.
- 19 Nyul L G, Udupa J K. On standardizing the MR image intensity scale [J]. Magnetic Resonance in Medicine, 1999, **42**(6): 1072-1081.
- 20 Worth A. The Internet Brain Segmentation Repository (IBSR) [EB/OL]. <http://www.cma.mgh.harvard.edu/ibsr/>, 2006-6-20/2007-6-15.
- 21 Horsch A. EFMI: European Federation for Medical Informatics-Working Group on Medical Image Processing [EB/OL]. <http://www.efmi-wg-mip.net/>, 2006-8-28/2007-6-16.
- 22 Yoo T. National Library of Medicine Insight Segmentation and Registration Toolkit (ITK) [EB/OL]. <http://www.itk.org>, 2006-11-13/2007-6-10.