

第5章

分割评价比较



教学视频

对图像分割的研究已进行了 60 多年,提出的方法成千上万,但尚没有一种适用于所有图像的通用算法。绝大多数算法都是针对具体问题提出的,这里一个重要的原因就是尚无通用的分割理论。另外,给定一个实际应用后选择适合的分割算法仍是一个很复杂的问题,且没有标准的方法。由于缺少通用的理论指导,要分割的图像又各有特点,进行分割常常需要反复进行试验。这些问题的存在促使人们一直不断地对图像分割开展大量工作[章 2014c]、[Zhang 2015e]、[章 2023]。

要克服上述图像分割中的问题和困难需要对如何评价(评估)图像分割技术及其性能开展研究。分割评价通过对分割算法性能的研究以达到改进和提高现有算法的性能、优化分割过程、改善分割质量以及指导新算法研究的目的。事实上,对分割算法的性能评价和比较[Zhang 1996a]也长期得到广泛的重视[Zhang 2015c]。

对图像分割的研究可分为 3 个层次。如果说对图像的分割处在研究的第一个层次,那么对图像分割的评价则构成研究的第二个层次(它帮助把握不同分割算法的性能),而对评价方法和评价准则的系统比较和刻画则构成研究的第三个层次(它帮助把握不同评价方法的性能)。换句话说,分割评价是为了研究分割技术,而对分割评价方法的比较和刻画则是为了研究评价方法,以更好地评价分割技术。从某种意义上说,对分割评价的比较和刻画是对分割评价的评价。

总体来说,在分割研究的 3 个层次都有许多值得进一步深入研究的地方。对图像分割技术自身理论的研究仍有待加强,对分割技术的评价研究仍不全面,而对评价方法的系统研究仍很缺乏。值得指出的是,研究图像分割评价正是为了更好地研究图像分割本身,而对评价方法的比较和刻画正是为了更好地研究分割评价,以便对分割研究本身起到引导作用。鉴于这项工作的重要性,本章将对图像分割评价已有的研究成果进行综述和讨论,以推动这方面的深入研究,并从“上”而“下”促进图像分割技术的进展。

根据上述讨论,本章各节将安排如下。

5.1 节首先对已有的评价研究进行分类,包括评价工作(分成两种)和评价方法(分成 3 组),为进一步的讨论打下基础。

5.2 节结合一个评价框架来介绍试验评价的机理和各个具体步骤,这是评价研究的重点之一。

5.3 节对三组评价方法中使用的各种评价准则进行详细介绍和分析,讨论它们各自的特点,这是评价研究中的另一个重点。

5.4 节给出了一个对各类分割算法进行评价的实例,介绍实验的设置和环境(包括涉及的算法和图像)以及得到的结果和对结果的讨论。

5.5 节将对图像分割的研究从评价提高到对评价的评价,通过对评价方法的特性分析以及对定量准则的实验比较,给出它们评价能力的优劣次序。

5.6 节介绍了近年若干在分割评价的比较和刻画中的相关工作。

5.7 节介绍了一个基于评价的分割算法优选系统,给出了一个利用分割评价的成果指导分割的例子。主要介绍了对算法优选的思想和策略以及优选系统的实现和效果。

5.1 分割评价研究分类

对图像分割评价的工作可以分成两种情况。

(1) **性能刻画**: 掌握某种算法在不同分割情况中的表现,通过选择算法参数来适应分割具有不同内容的图像和分割在不同条件下采集到的图像的需要。

(2) **性能比较**: 比较不同算法在分割给定图像时的性能,以帮助在具体分割应用中选取合适的算法或改进已有的算法。

上述这两方面的内容是互相关联的,性能刻画能使对算法的性能比较更加全面,性能比较能使对算法的性能刻画更有目的性。

为达到分割评价的目的,对评价方法提出的基本要求主要有[Zhang 1994]:

(1) 应具有广泛的通用性,即评价方法要适于评价不同类型的分割算法并适合各种应用领域;

(2) 应采用定量的和客观的性能评价准则,这里定量是指可以精确地描述算法的性能,客观是指评判摆脱了人为的因素;

(3) 应选取通用的图像进行测试以使评价结果具有可比性和可移植性,同时这些图像应尽可能反映客观世界的真实情况和实际应用领域的共同特点。

现有评价分割技术和算法的方法可归纳为两大类:一类是直接的方法,它直接研究分割算法本身的原理特性,通过分析推理获得分割算法的性能,所以也称为**分析法**;另一类是间接的方法,它根据对试验图像的分割质量间接地评判分割算法的性能,所以也称为**试验法**。具体就是使用需要评价的算法去分割图像,然后借助一定的质量测度来判断分割结果的优劣,据此转而出所得分割算法的性能。试验法可进一步分为两组:一组采用(常根据人的直觉建立的)一些**优度**参数来描述已分割图的特征,然后根据参数值来判定分割算法的性能;另一组则先确定理想的或期望的分割结果参考图,然后通过比较已分割图与参考图之间的差异值来判定分割算法的性能。前一组方法可称为**优度试验法**,后一组方法可称为**差异试验法**。综合以上讨论,分割评价方法可分为分析法、优度试验法和差异试验法 3 组,它们各自的特点及相互关系可借助图 5.1.1 来说明[Zhang 1996a]。

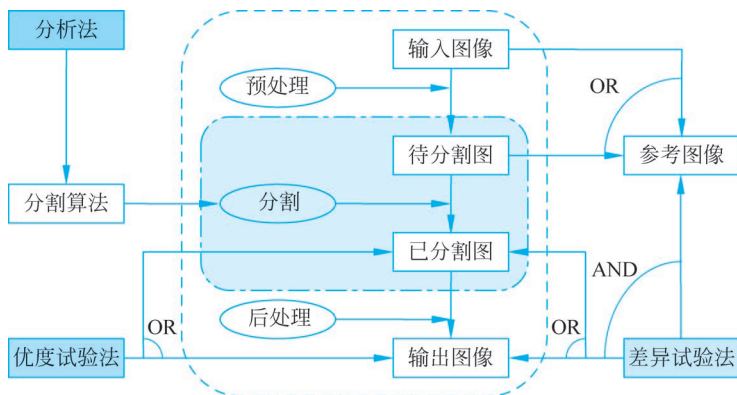


图 5.1.1 图像分割评价方法分组

图 5.1.1 中点画线框内给出了一个狭义的图像分割基本流程图,这里将图像分割看作用

分割算法去分割待分割图像得到已分割图像的过程。图 5.1.1 中虚线框内给出了一个广义的图像分割基本流程图,这里将图像分割看作由 3 个步骤串联而成:一是预处理;二是狭义的图像分割;三是后处理。在广义的分割流程图中,待分割图像是通过一般的输入图像进行一定的预处理后得到的;而已分割图像也还要经过一定的后处理才能成为最终的输出图像。

从图 5.1.1 可以看到上述 3 组评价方法的不同作用点和工作方式。分析法仅仅作用在分割算法本身上,并不涉及分割图像和分割流程。优度试验法是去检测已分割图像或输出图像的质量来评价分割算法,不必考虑分割流程的输入;而差异试验法则是将通过输入图像或待分割图像得到的参考图像与已分割图像或输出图像的质量进行对比来评价分割算法,所以既要考虑分割流程的输入,又要考虑分割流程的输出。

不管是分析法还是试验法(包括直接法或间接法),它们基本上都有两个关键步骤或内容:一个是对分割算法进行分析或试验的框架或机制以及途径或方案(即机理和程序);另一个是用来评判算法特性的评价准则(也常称测度或指标)。前者对同一组评价方法有类似之处(因为这里分组时主要考虑的是机理),而后者则在同一组方法中对各个具体的方案也各有特点。评价准则对分析法或试验法都很重要。对分析法来说,要分析算法的某些特性,需要有衡量这种特性的测度,算法的特性要根据这种测度才能判别和比较。对试验法来说,因为它们依靠对图像分割后的结果进行评判来决定所用分割算法的性能,因而定量的(可计算和比较的)和客观的(不会因为评价人主观意识而不同)评价指标至关重要,算法的优劣就是基于指标值来确定的。由此可见,要对图像分割进行评价必须要采用合适有效的性能评价准则。另外,评价准则在一定情况下也对评价方案有限定作用,一定的评价准则只能用在一定的评价方案中。

5.2 分割算法评价框架

评价机制主要体现在评价框架中。其中,差异试验法既要考虑分割算法本身又要考虑分割流程,在分割流程中还需同时考虑输入和输出,相对比较复杂。

图 5.2.1 给出了一个对分割算法进行差异试验评价的框架示意图,它主要包括 3 个模块:性能评判、图像合成、算法测试[Zhang 1992b]。

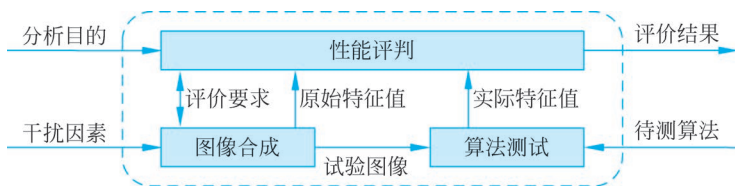


图 5.2.1 分割算法评价框架

如图 5.2.1 所示的评价框架是通用的。一方面,有关分析目的、评价要求、图像获取及处理的条件和因素可以被有选择地结合进这个框架,因此它可以适用于各种应用领域。另一方面,因为它在研究分割算法时只需要用到图像分割的结果而不需要了解被研究算法的内部结构特性,所以可适用于所有的分割算法。下面对这 3 个模块分别给予简单介绍。

1. 性能评判

性能评判包括 3 个相关联的部分(见图 5.2.2)。

(1) 特征选取:根据分割目的来选取相应的目标特征以进行评价,同时相应的合成测试图也要据此产生。

(2) 差异计算:利用从原始图和分割图所得到的原始和实测特征值计算。

(3) 性能描述: 将差异计算结果与图像合成条件结合以给出评价结果。

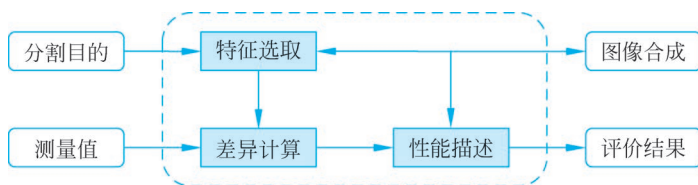


图 5.2.2 性能评判模块示意图

2. 图像合成

为根据试验分割结果来评判分割算法需要采用合适的分割试验图。为了保证评价研究的客观性和通用性,可采用合成图像来测试分割算法并作为参考分割图。这样不仅客观性好,而且可重复性强,结果稳定。而如用真实图则研究结果常受限于具体的应用,并且由于需要人工分割以得到参考图从而会在评价中引入主观偏差。合成图像应用中重要的一点是生成的图像应能尽可能反映客观世界,这就需要结合应用领域的知识。**图像合成**流程应可以调整以适应诸如图像内容变化、各种获取图像的条件等实际情况。下面介绍的图像合成系统框架可以满足以上的要求,它包括 4 个相关的部分(见图 5.2.3)。

(1) 组建基本图: 基本图反映基本内容和结构,可以根据实际应用领域的模型来建立,它将是生成一系列合成图的基础和起点。

(2) 目标调整: 修改基本图中的目标以产生不同灰度、尺寸、形状、数量、位置等的目标,以模拟实际图像;有关目标的原始数据可输出给性能评价模块。

(3) 叠加干扰: 通过模拟采集条件(也可实际采集)产生噪声,模拟加工情况(如平滑)产生模糊,并叠加到具有不同目标的图像中,从而逼近真实世界。

(4) 图像组合: 按一定次序组合各种图像目标和干扰因素以最终获得尽可能接近于实际情况的试验图像以用于图 5.2.1 中的“算法测试”模块。

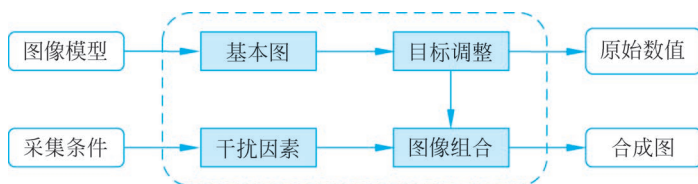


图 5.2.3 图像合成模块示意图

例 5.2.1 用于分割评价的合成图示例

图 5.2.4 给出了一组根据如图 5.2.3 所示的流程合成的分割评价试验图,这些图均为 256×256 px, 256 级灰度图。基本图是将亮的圆形目标放在暗的背景正中而组成的。图中目

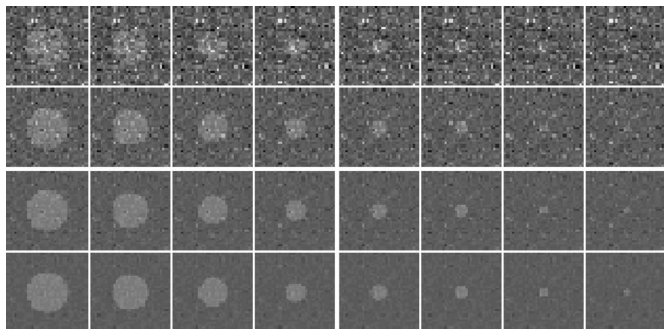


图 5.2.4 分割评价试验图示例

□

标与背景间的灰度对比度为 32, 叠加的噪声均为零均值高斯随机噪声。这组图可称为“尺寸组”, 从左至右 8 列图中的目标面积分别为全图的 20%、15%、10%、5%、3%、2%、1%、0.5%, 从上至下 4 行图的信噪比分别为 1、4、16、64。

3. 算法测试

算法测试的流程框图见图 5.2.5。

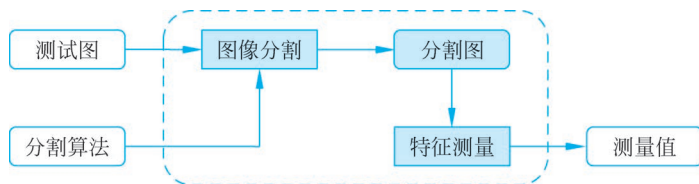


图 5.2.5 算法测试的流程框图

如图 5.2.5 所示的算法测试流程框图是一个典型的图像分析模块框图, 它包括两个前后连接的步骤: 分割和测量。在分割阶段, 将被测算法看作一个“黑盒子”, 对它的输入是测试图(试验图), 而得到的输出是分割图。在测量阶段, 根据分割出来的目标对预先确定的特征进行测量就得到实际的目标特征值, 然后将这些特征值输入图 5.2.1 中的“性能评判”模块以进行差异计算。

5.3 分割评价的准则

在 5.1 节中, 将分割评价方法分成了 3 组: 分析法、优度试验法和差异试验法。不同方法所采用的评判算法特性的评价准则也各有特点。下面分组介绍一些基本和典型的评价准则。

5.3.1 分析法准则

分析法准则指适合于分析分割算法本身的评价准则, 可以是定性的也可以是定量的。

1. A-1: 所结合的先验信息

图像分割属于图像工程中层的图像分析, 本身已有一定的抽象程度, 而高层知识的指导作用也比较重要。在设计分割算法时, 一些有关所要分割图像自身的特性信息可以被结合和使用, 以提高分割的稳定性、可靠性和效率等。对某些算法来说, 将实际应用中的先验信息或先验知识结合进去有可能取得更好的分割结果。由于先验信息的应用对分割算法的性能有较大的影响, 所以根据不同算法所结合的先验信息的种类和数量的不同, 可以从一定程度去比较算法的优劣。

需要注意的是, 并不是对所有算法都可有效地使用这种准则来评判比较。一方面这是因为目前还没有可以完全定量描述先验信息的方法, 因而难以比较。另一方面, 不仅是先验信息的种类和数量, 而且如何在分割算法中利用这些信息也会对算法的性能有很大影响[Zhang 1991c]。由于先验信息的利用程度是很难定量描述的, 所以这种准则主要用于定性分析算法性能。

2. A-2: 处理策略

图像分割和许多图像分析技术一样, 其过程可以串行或并行、迭代或混合地实现。图像分割算法的性能常常与这些处理策略是紧密联系的, 所以根据算法的处理策略也可在一定程度上把握算法的特性[Zhang 1993a]。例如, 并行工作的算法的处理速度较高, 尤其适合于用具有并行处理能力的计算机来快速实现。但是由于在并行处理中对所有像素是同时处理的, 无

法利用串行和迭代算法中的中间结果调整算法,所以不够灵活,受噪声和干扰的影响也会较大。反过来,串行工作的算法虽然比较复杂,速度较低,但由于较多地利用了前期逐步获取的信息,能较好地应对困难的情况,抗噪声能力也常较强。

3. A-3: 计算费用

每个分割算法都由一系列的运算操作来实现。为完成这些操作所需的**计算费用**与分割过程的复杂性、算法的效率(以及计算速度)都有关,也是衡量算法性能的一个指标。实际中完成各种操作的计算费用还与许多因素,例如计算机硬件、图像内容等都有关。为消除计算机硬件的影响,计算费用也可按不同操作的类型和数量来计算[Zhang 1993a]。在这种情况下,对某些算法的计算费用可由对算法的分析定量得到。有些分割算法的计算量是与图像内容或要分割的区域有关的,或者说与图像本身的复杂度有关,这时就需要用到根据每个具体分割任务来计算的准则(即通过测量具体分割图像所需的时间)来判定算法的计算费用,不过这一般仅对某些算法适用。

4. A-4: 检测概率比

检测概率比被定义为正确检测概率与错误检测概率之比。这个准则最初用来比较和研究各种边缘检测算子。给定一个边缘模式,一个算子在检测这类边缘时的正确检测概率 P_c 与错误检测概率 P_e 可以由下两式计算出来(T 为一个给定的阈值):

$$P_c = \int_T^{\infty} P(t \mid \text{edge}) dt \quad (5.3.1)$$

$$P_e = \int_T^{\infty} P(t \mid \text{no-edge}) dt \quad (5.3.2)$$

对简单的边缘检测算子常可以通过分析得到其 P_c 和 P_e 的比值。这个值越大,表明算子在检测对应边缘时可靠性越高。由于许多分割技术利用边缘检测算子来帮助分割图像,所以这种准则可用来对这些算法的性能进行评价。实际中,常将 P_c 和 P_e 的比值用检测和误警曲线表示以评估分割性能。

5. A-5: 分辨率

利用不同的分割算法所得到的分割图像其**分辨率**可以有多种情况,例如,可以是一个像素、若干像素(组合)或一个像素的若干分之一(亚像素或子像素)。实际中大多数算法的分辨率都以像素为单位,不过纹理图像的分割常以多个像素的集合为单位,而**亚像素**的分辨率在很多应用中得到关注(见 4.3 节),所以分辨率也是衡量算法性能的一个有效指标。

一般来说,对给定的分割算法,其分辨率通过分析其原理和步骤就可得到。

5.3.2 优度试验法准则

优度试验法在评价图像分割效果时要利用一些**优度准则**,这些准则常代表主观上对理想分割结果所期望的一些性质。优度准则给出的结果常被量化,一般是定量的。

1. G-1: 区域间对比度

图像分割要把一幅原始图像划分成若干区域。直观地考虑,这些区域的特性之间应有比较大的差距,或者说有明显的对比。根据对给定特性的**区域间对比度**的大小可以判别分割图的质量,也可由此推出所用分割算法的优劣。对图像中相邻接的两个区域来说,如果它们各自的平均灰度为 f_1 和 f_2 ,则它们之间的**灰度对比度**(GC)可按下式计算[Levine 1985]:

$$GC = \frac{|f_1 - f_2|}{f_1 + f_2} \quad (5.3.3)$$

事实上,式(5.3.3)中的 f_1 和 f_2 也可代表除灰度外的其他特征量。这样就得到其他特征的区域间对比度。当一幅图中有多个区域时,可利用式(5.3.3)分别计算两两邻接区域之间的对比度再组合起来。

与区域间对比度密切相关的一个指标被称为**相关性**[Brink 1989],虽然它直接测量的是原始图像和阈值化后的二值图像之间的相关系数。不过已经证明,这里相关系数的平方也就是一种熵分割算法[Otsu 1979]中的类分离熵,所以这个指标的含义与区域间对比度类似[Zhang 1996a]。

2. G-2: 区域内均匀性

分割常被定义为要把一幅原始图像 $f(x, y)$ 分解成若干内部具有相似特性的区域,所以可用分割图中各区域内特性均匀的程度来描述分割图像的质量,这里使用的性质就是**区域内均匀性**。如以 R_i 表示分割图中的第 i 个区域, A_i 表示其面积,则分割图中各区域内的**均匀性测度**(UM)可表示为[Sahoo 1988]

$$UM = 1 - \frac{1}{C} \sum_i \left\{ \sum_{(x,y) \in R_i} \left[f(x,y) - \frac{1}{A_i} \sum_{(x,y) \in R_i} f(x,y) \right]^2 \right\} \quad (5.3.4)$$

其中, C 为归一化系数。与此类似,还有**繁忙性测度**[Weszka 1978]和**高阶局部熵**[Pal 1993]等测度也是基于相同想法的。

3. G-3: 形状测度

直观地讲,一个好的分割结果应满足某些主观条件或视觉要求。一般希望物体轮廓线比较平滑,**形状测度**就是用于衡量目标外轮廓的光滑程度的,所以也可称轮廓指标。如以 $f_N(x, y)$ 表示像素 (x, y) 的邻域 $N(x, y)$ 中的平均灰度, $g(x, y)$ 表示像素 (x, y) 处的梯度,对图像以 T 为阈值进行分割所得形状测度 SM 可用下式计算[Sahoo 1988]:

$$SM = \frac{1}{C} \left\{ \sum_{x,y} \text{Sgn}[f(x,y) - f_{N(x,y)}] g(x,y) \text{Sgn}[f(x,y) - T] \right\} \quad (5.3.5)$$

其中, C 为归一化系数; $\text{Sgn}(\cdot)$ 代表单位阶跃函数; T 是预先确定的阈值。

5.3.3 差异试验法准则

差异试验法中所使用的评价准则用来比较已分割图与参考图之间的差异,所以称为**差异准则**。利用差异准则应能定量地衡量分割结果,且这种量度是客观的。

1. D-1: 像素距离误差

实际中的分割结果常常不是完善的,在这种情况下总有一些像素被错误地划分到并不应属于的区域。这些被错分的像素与它们本应该属于的正确区域之间的距离(带有一定的空间信息),从一个角度反映了分割质量的好坏。现已提出了若干基于**像素距离误差**的测度来评价分割结果。一个常用的测度是**质量因数/品质因数**(FOM):

$$FOM = \frac{1}{N} \sum_{i=1}^N \frac{1}{1 + p \times d^2(i)} \quad (5.3.6)$$

其中, N 是错分像素的个数; p 是一个比例系数; $d^2(i)$ 代表第 i 个错分像素与其正确位置的距离。与此相关联的一个测度是**偏差的平均绝对值**(MAVD):

$$MAVD = \frac{1}{N} \sum_{i=1}^N |d(i)| \quad (5.3.7)$$

还有一个测度称为**归一化距离测度**(NDM):

$$NDM = \frac{\sqrt{\sum_{i=1}^N d^2(i)}}{A} \times 100\% \quad (5.3.8)$$

式中, N 和 $d^2(i)$ 同式(5.3.6); A 是图像的面积。其他类似的测度还有像素空间分布、修正的质量因数(MFOM)、概率加权的质量因数(PWFOM)等。

2. D-2: 像素数量误差

对图像分割结果来说,由于分割错误而产生的错分像素个数是一个重要的图像质量衡量指标。对此已提出了借助像素数量误差来评判分割图像的质量的许多不同加权方法。例如,有一种称为误差概率(PE),当图像是由目标和背景两部分构成时,PE可用下式计算:

$$PE = P(o) \times P(b | o) + P(b) \times P(o | b) \quad (5.3.9)$$

其中, $P(b|o)$ 是将目标错误划分为背景的概率; $P(o|b)$ 是将背景错误划分为目标的概率; $P(o)$ 和 $P(b)$ 分别是图像中目标和背景所占比例的先验概率。当图像中包含多个目标时,一个更加通用的 PE 定义可见[Lim 1990]。误差概率还有一个变型称为误分率[罗 1997]。

对一些特定的分割算法,其使用的分割参数与分割后的像素数量误差有一定的对应关系,此时像素数量误差也可借助参数误差来计算。例如,对全局阈值化的算法,当阈值给定时,像素数量误差也确定了,所以根据某一算法实际得到的阈值和理想阈值之间的差也可衡量这一算法的优劣。

3. D-3: 目标计数一致性

设 S_n 为对一幅图像进行分割所得到的目标个数, T_n 为图像中实际存在的目标个数,由于分割结果的不完善,所以 S_n 和 T_n 有可能不同。它们之间的差异在一定程度上反映了分割算法性能的一个方面。利用概率的方法可借助这个差异直接定义一个称为目标计数一致性(OCA)的测度来评价分割算法。还有一种借助这个差异描述算法性能的指标称为图像分块数,用 F 表示,其定义为

$$F = \frac{1}{1 + p |T_n - S_n|^q} \quad (5.3.10)$$

其中, p 和 q 均是尺度参数。

4. D-4: 最终测量精度

图像分析中的一个基本问题就是要获得对图像中各个目标特征值的精确测量[Young 1993],这是图像分析中进行分割和其他后续操作的最终目标。因为特征的测量是基于分割结果的,所以其精确度直接取决于分割的结果和分割算法的性能。另外,这个精度(可称为最终测量精度,UMA)也反映了分割图像的质量并可以用来评判算法的性能[Zhang 1994]。从高层图像理解的角度看,一幅分割图像的质量高低要看基于它做出的决策和基于原始图像做出的决策有多大区别,这也是最终测量精度的本意所在。

在实际中,为了描述目标的不同性质可以使用不同的目标特征,因此 UMA 可写成 UMA_f 。这给出了以目标特征为参数的一系列评价准则。如果用 R_f 代表从参考图像中获得的原始特征量值,而 S_f 代表从分割后图像中获得的实际特征量值,则它们的绝对差和相对差可分别由以下两式算得:

$$AUMA_f = |R_f - S_f| \quad (5.3.11)$$

$$RUMA_f = \frac{|R_f - S_f|}{R_f} \times 100\% \quad (5.3.12)$$

由上面两式分别得到绝对最终测量精度和相对最终测量精度。注意, $AUMA_f$ 和 $RUMA_f$ 的值都反比于分割质量: 它们的值越小,说明分割效果越好,所用的算法性能越好[章 1997b]、[Zhang 1997b]。

最终测量精度准则能满足 5.1 节中所述进行评价的 3 个要求。首先它们面向分析目的,

因而直接反映了人们对分析工作的质量量度的需要。而且因为这是不同类型分割算法的共性且不依赖于分割算法及其应用领域,所以准则具有广泛的通用性。其次它们是客观的,因为它们反映了目标本身的特性而不是主观定义的视觉质量;而且它们又是定量的,因为根据式(5.3.11)和式(5.3.12)可以精确地计算它们的数值。最后,好的评价方法应选取通用的图像进行测试以使评价结果具有可比性和可移植性,而最终测量精度准则计算简单,很易从通用的图像中算出。

例 5.3.1 用于最终测量精度的几个特征的比较

各种特征不仅对目标的描述能力不同,而且用于 UMA 中时其分割评价的性能也会不一样。图 5.3.1 给出了对 5 个常用的特征[即目标的面积(A ,见 7.2.1 小节)、圆形性(C ,见 11.4.1 小节)、偏心率(E ,见 11.4.1 小节)、形状因子(F ,见 11.4.1 小节)和周长(P ,见 7.1.1 小节)]进行试验得到的一些结果。图像分割算法的性能可由算法的参数控制,所以改变算法参数可得到一系列效果不同的实际分割图。图 5.3.1 中横轴对应算法参数的变化,它们对应分割效果从差到好又从好到差的变化,纵轴对应被归一化到 $[0,1]$ 区间的 UMA 值,值大表示分割效果差,值小表示分割效果好。

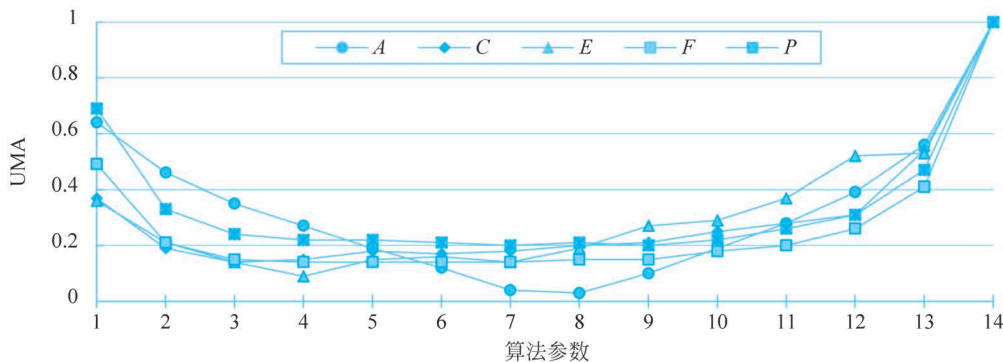


图 5.3.1 特征性能研究结果

从图 5.3.1 中可看到,各个特征的 UMA 值都随算法参数的改变而呈现从大到小又从小到大的变化,这说明这些特征用在 UMA 中都有表示不同分割质量从而评判分割算法性能的能力。不过这些 UMA 曲线也有差别,这反映了各对应特征评判分割的能力不同。具体来说,可从两方面进行分析,一方面是看曲线的全局形状特别是谷的深度。因为这里特征值已归一化到了相同的范围,所以谷的深度反映了各个特征 UMA 值的动态范围,谷的深度越大,说明对不同分割结果的区别能力越强。另一方面是看曲线的局部光滑程度。它反映了特征 UMA 值稳定跟踪分割结果微小变化的情况。一般说来,曲线平滑表明特征反映相近分割结果比较一致和稳定。根据以上标准考察图 5.3.1,可以发现各曲线的不同特点,其中圆形性和偏心率的曲线不够光滑而形状因子和周长曲线在某些范围内几乎是水平的。相比之下,面积曲线具有较明显的谷和较平滑的升降段,这表明在使用 UMA 的评价中,面积这个特征比其他几个特征常能较好地判断分割结果和质量。□

5.4 分割算法评价实例

借助前面介绍的分割算法评价框架和评价准则,可对不同的分割算法进行评价。下面介绍一个实际评价工作的部分实验和结果[Zhang 1997a]。

5.4.1 实验算法和图像

根据图 5.2.1 给出的对分割算法进行试验评价的框架,先介绍用于评价的算法和图像。

1. 所评价的算法

根据 2.1 节中对分割算法的分类,为使评价具有较强的通用性和代表性,从每一类算法中各选了一种典型的算法[章 2005b]。

- (1) 算法甲(PB类): 坎尼算子边缘检测及边界闭合法,参见 2.2.4 小节和 2.2.5 小节。
- (2) 算法乙(SB类): 动态规划轮廓搜索法,参见 2.3 节和[章 2005b]。
- (3) 算法丙(PR类): 改进的直方图凹凸性分析法,参见 2.4.2 小节。
- (4) 算法丁(SR类): 分裂、合并和组合法,参见 2.5.2 小节。

2. 所用的实验图

在研究中采用了两组实验图:一组实验图即例 5.2.1 中的目标尺寸组(可用来研究目标尺寸与分割效果之间的关系,见下面对算法丁的讨论);另一组实验图称为目标形状组(可用来研究目标形状与分割效果之间的关系,见下面对算法乙的讨论)。两组图合起来如图 5.4.1 所示,其中,左 8 列属于目标尺寸组(从左向右分别为 1 号到 8 号圆形目标),右 4 列属于目标形状组(从左向右分别为 1 号到 4 号椭圆目标)。这两组图有一列图是相同的,即左数第 8 列图。在图 5.4.1 中,目标形状组里沿水平方向改变目标形状,沿竖直方向改变图像信噪比。目标形状的改变是靠改变基本图中的圆形目标为不同的椭圆来实现的,从左到右 4 列图中目标的长短轴比分别为 $1:1$ 、 $1.5:1$ 、 $2:1$ 、 $2.5:1$ 。为了消除目标尺寸对分割的影响,在改变目标形状时要保持目标的面积与基本图中的目标尽量一致[Zhang 1998a]。噪声叠加采用了与例 5.2.1 中相同的方法进行。

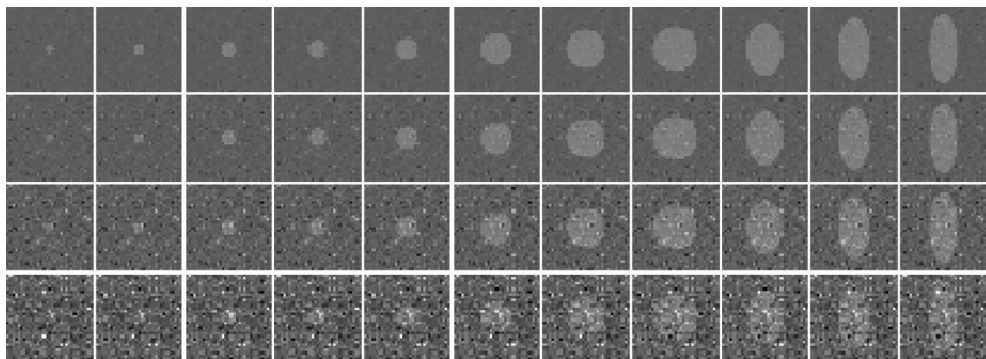


图 5.4.1 目标尺寸组和目标形状组实验图

5.4.2 实验结果和讨论

下面仅介绍对两种串行方法,即算法乙和算法丁得到的实验结果,并进行相应的讨论。

1. 对算法乙的讨论

用算法乙分割目标形状组的图像,所得到的一些结果如图 5.4.2 所示。图 5.4.2 中的上下两行图像分别对应含有 1 号椭圆目标和 4 号椭圆目标的各种情况,其中从左到右的 6 列图像分别为原始图像,对无噪声图像分割的结果,以及对有噪声且其信噪比分别为 64、16、4、1 的图像的分割结果。

以目标形状为参数,相对最终测量精度作为信噪比函数的曲线如图 5.4.3(a)所示。在图 5.4.3(a)中,除对应 4 号椭圆目标图像的曲线外,其他几条曲线均比较类似。注意,这里尽

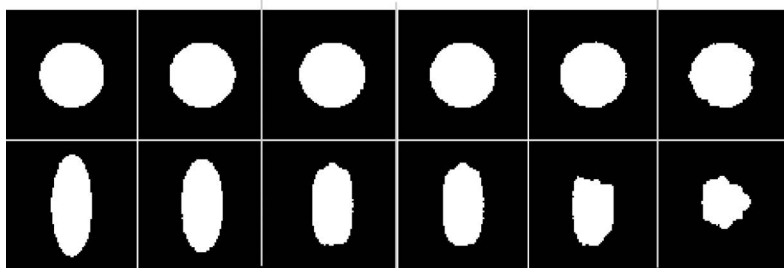


图 5.4.2 用算法乙得到的分割结果

管各个目标的形状不同,但各实验所用的感兴趣区域(这里是包含椭圆轮廓的圆环)是相同的,所以这些曲线的类似说明分割结果与感兴趣区域和目标之间的相对形状关系基本无关。换句话说,对算法乙来讲,确定感兴趣区域在这里并不是影响分割的重要因素。而信噪比对分割的影响要明显得多,正如图 5.4.3(a)中曲线随信噪比快速下降所表明的那样。

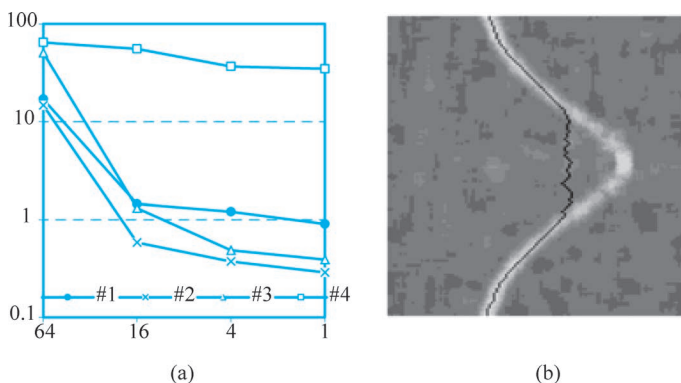


图 5.4.3 算法乙受目标形状的影响

回过头来再讨论对应 4 号椭圆目标图像的曲线,这时利用极坐标变换后的图像来讨论比较清晰。图 5.4.3(b)给出了用梯度算子运算并进行极坐标变换得到的结果,其中竖轴对应极角,横轴对应矢径(从感兴趣区域中心到目标边界)。图 5.4.3(b)中的浅色条带基本对应目标的边界,但利用动态规划搜索技术自上而下搜索得到的边界线如图 5.4.3(b)中的黑线所示,有相当一部分脱离了预期的范围。对这个问题经进一步分析,发现主要是由于在极坐标变换时采样密度不够造成的,对细长或非规则目标需要适当增加采样率以避免这个问题[Zhang 1993d]。

2. 对算法丁的讨论

下面讨论算法丁的实验情况。这里使用了目标尺寸组的实验图。在分裂合并算法中,用于判断应该进行分裂或合并的一致性准则很重要。一种常用的准则基于区域的方差。由于区域的方差会随噪声而变化,所以算法丁的准则参数 V 要根据图像中的噪声水平来选择。如果噪声不知,就会导致准则参数 V 选择不准确,这会给分割造成什么影响呢?为此,考虑做出算法丁的以信噪比为参数,实际获得的相对最终测量精度 RUMA 与理想的相对最终测量精度 RUMA 的比值(纵轴)作为准则参数与噪声标准方差(SDN)的比值(横轴)的函数曲线,见图 5.4.4,其中,图 5.4.4(a)和图 5.4.4(b)分别为用 8 号圆形目标和 5 号圆形目标得到的结果。这里当函数值大于 6 时都标为 6,以使其他数值显示得比较清晰。

由于信噪比 SNR 和噪声标准方差 SDN 是相关的,所以对每个 SDN,选了 5 个 V 的值,使得 V 与 SDN 的比值分别为 0.5、0.75、1、1.25 和 1.5。实际的相对最终测量精度 RUMA 是在不同的 V 与 SDN 的比值下测得的,而理想的相对最终测量精度 RUMA 则是在 V 与 SDN 的比值为 1 时(即准确设置 V 参数的情况)测得的。实际 RUMA 和理想 RUMA 的比值越大,表示分割结果越差。

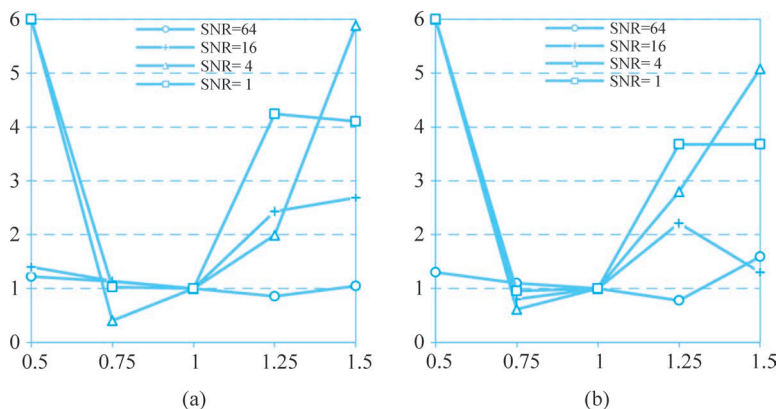


图 5.4.4 对算法丁的评价结果

图 5.4.4(a)和图 5.4.4(b)有许多相似之处,当 V 与 SDN 的比值不为 1 时,大多数实际的 RUMA 与理想的 RUMA 的比值都大于 1。当 V 与 SDN 的比值大于 1 时,分割算法趋向于欠分割图像,而当 V 与 SDN 的比值小于 1 时,分割算法趋向于过分割图像。图 5.4.4(a)和图 5.4.4(b)中部的低谷表明,将 V 设得比 SDN 大或将 V 设得比 SDN 小都会影响分割算法的性能。另外应指出,过分割图像所需要的计算量相比欠分割图像所需要的计算量会多得多 [Zhang 1997c]。

图 5.4.4(a)和图 5.4.4(b)也有不同之处,最明显的就是在图 5.4.4(a)中只有在 SNR=4 和 SNR=1 两种条件下,实际的 RUMA 与理想的 RUMA 的比值当 V 与 SDN 的比值为 0.5 时达到最大值 6;而在图 5.4.4(b)中,在 SNR=16 条件下,当 V 与 SDN 的比值为 0.5 时也达到最大值 6。这个区别表明不正确的设定 V 在图像目标尺寸较小时影响更大,即在分割小目标时 V 的设定更为关键。

图 5.4.5 给出了当使用不同的 V 与 SDN 比值时分割同一幅图得到的结果。上下两行分别对应都使用 8 号圆目标,但图像信噪比分别为 SNR=16 和 SNR=1 的两种情况。从左向右, V 与 SDN 的比值分别为 0.5、0.75、1、1.25 和 1.5。由图 5.4.5 可见, V 与 SDN 的比值较小时(即过分割时),分割结果细节较多;而 V 与 SDN 的比值较大时(即欠分割时),分割结果的轮廓比较规则但粗糙。

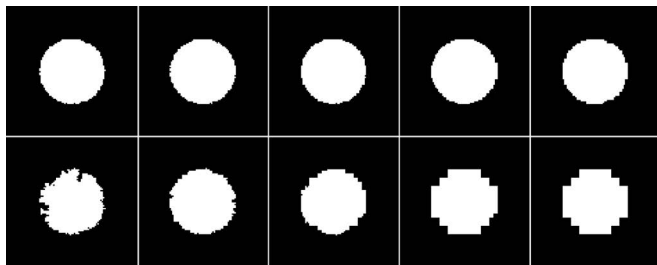


图 5.4.5 用算法丁得到的分割结果

以上实验结果和讨论对在实际应用中有效选择和使用合适的分割算法时可提供依据。

5.5 评价方法和准则比较

前面几节介绍了多种已提出的主要分割评价方法和基本的评价准则,这些方法和准则采用的原理和机理各有特点,因而评价性能也会不同。本节讨论对分割评价方法和评价准则的

性能比较研究。事实上,对分割评价方法的系统比较和刻画对选取有效的评价方法至关重要,它对评价分割算法和提高分割质量也非常关键。

本章开头提到,如果将研究图像分割技术作为图像分割研究中第一个层次的内容,则研究图像分割评价技术可看作图像分割研究中第二个层次的内容。那么如何评价这第二个层次的图像分割评价技术呢?这就需要第三个层次的研究,即对图像分割评价技术进行系统的性能刻画和比较。

在图像分割第三个层次的研究中也可采用在第二个层次研究中的一些概念和思路。例如,对评价方法和准则的比较也可采用分析或试验的方法,分析法直接研究评价方法或准则本身,讨论它们的原理、目标、应用性、代价等;试验法则将评价方法或准则应用于评价工作,根据它们在评价过程中的表现来确定它们的优劣。

下面先对评价方法进行讨论和对比,然后对各种评价准则进行分析比较,最后对其中几个典型的定量试验准则(包括优度试验准则和差异试验准则)进行实验比较。

5.5.1 方法讨论和对比

对 5.1 节的 3 组评价方法可从 4 方面进行分析讨论和对比。

1. 通用性

现已提出的分割算法的种类很多,所以用来评价它们的方法和准则应具有通用性,即应当可以适用于研究各种不同类型的分割算法。有些方法和准则由于自身的限制只适合于研究某些特定的分割算法,这样就会使评价工作受到限制。例如,对 5.3.2 小节中形状测度的计算需要用到阈值化算法中的阈值,所以它只可用于评价阈值化算法。另外,如 5.3.1 小节中的检测概率比对复杂的算子很难得到解析表达式,也使它的应用受到一定限制。

2. 复杂性

一种评价方法能否实用与它本身实现的复杂性,或者说与为了进行评价所需的操作手段和工作量有很大关系,这点对评价方法能否在线应用也非常重要。有些准则对每个像素都需要进行一系列的计算(如 5.3.2 小节中对区域内均匀性测度的计算),而一幅图像中的像素数量很大,所以评价所需的计算量就会很大。还有些试验指标,为了算得它们的数值除了需要对图像进行分割以外还需要其他额外的处理和辅助运算(如 5.3.3 小节中介绍的像素距离误差),这样评价工作就会变得相当复杂或需要很大的计算量。

3. 主客观性

对每种评价方法来说,其背后常有某些主观或客观的考虑,或者说所采用的评价准则是根据特定的主观或客观因素而确定的。基于主观因素的准则有可能与人的直觉相吻合,但并不一定能够反映实际应用要求[Zhang 1994]。而基于客观因素的准则常可以提供一致的和无偏的研究结果[Young 1993]。图像分割作为图像分析中的第一个步骤,客观的结果更为重要。

4. 对参考图的需求

对有些评价方法,评价结论取决于将分割结果和参考(真值)图进行的比较,而参考图的获得对评价方法的实用性带来了一些特定的问题。对实际图像,精确的分割结果往往不能自动获得,手工分割也很难保证评价的一致性。由图 5.1.1 可知,分析法和优度试验法均不需参考图,而差异试验法则需要使用参考图。

除以上 4 个方面外,另外还有一些情况值得讨论。例如,由 3 组评价方法不同的作用点和作用方式可知:分析法评判算法可以只对算法本身进行分析,而并不需要实现算法本身(这样在评价时不会受实现算法时产生的误差或近似等影响);采用试验法则需要实现算法并对输

入图像进行实际分割以得到输出分割图(对差异试验法还需获得参考图)。又如分析法完全没有考虑算法的应用环境,评价结果只与算法本身有关。**优度试验法**实际上将已分割图的某些期望性指标结合在优度参数中,从而与实际应用建立了联系。**差异试验法**采用由输入图或待分割图获得的参考图作为分割的标准,已充分考虑了特定应用。在与应用结合方面,3组评价方法的差别也对它们的性能有很大影响。

5.5.2 准则的分析比较

以上讨论基本上以各组方法为单位,各组的准则各不相同。对参考图的要求根据准则所属组别即可确定,而对其他3方面的情况,需分别讨论。对5.3节介绍的各种评价准则进行分析可综合得到表5.5.1[Zhang 1993a],其中还补充了一些近年出现的评价准则(D-5~D-8)[Zhang 2015c]。从表5.5.1中可看出各组各个准则相互之间的一些主要优点和缺点。

表 5.5.1 分割评价准则的比较

组	评价准则	通用性	复杂性	主客观性	其他特点
A-1	先验知识	部分算法	低	主观	不同种类知识很难相互比较
A-2	处理策略	所有算法	低	客观	算法效率或复杂度的指示参数
A-3	计算费用	部分简单算法	低	客观	相对测度,未考虑硬件和软件实现情况
A-4	检测概率比	部分简单算法	中	客观	对复杂算法难以分析
A-5	分辨率	所有算法	低	客观	算法能力和效果的指示参数
G-1	区域间对比度	所有算法	中	主观	模拟人类评价能力的动态在线测度
G-2	区域内均匀性	所有算法	高	主观	与区域尺度和均匀性测度均有关
G-3	形状测度	阈值化算法	高	主观	仅与目标边界粗糙度有关
D-1	像素距离误差	所有算法	高	客观	需要与其他准则配合使用
D-2	像素数量误差	所有算法	中	客观	没有利用空间信息
D-3	目标计数一致性	所有算法	低	客观	简单但粗糙,当图像中目标数较少时不易精确
D-4	最终测量精度	所有算法	中	客观	一组与分割目的直接相关的测度
D-5	各种目标特性	所有算法	中	客观	一组表示分割目标特性的测度
D-6	区域一致性	部分算法	高	客观	与区域之间的相似性有关
D-7	灰度差别	所有算法	中	客观	平均灰度在分割前后的变化
D-8	对称散度(交叉熵)	所有算法	高	客观	反映图像之间的信息联系

进入21世纪,图像分割评价有了进一步的进展[Zhang 2006a]、[Zhang 2009c]、[Zhang 2015c]。表5.5.2给出了对近年出现的一些分割评价方法的分析比较。对每种方法列出了所用的主要评价准则(有些方法所用准则对原有准则有所改进,但基本上还属于同一类型)。有些方法结合使用了不止一种准则。所有准则均为**优度准则**或**差异准则**。

表 5.5.2 对21世纪一些新评价方法和准则的分析比较

编号	参考文献	主要准则	通用性	复杂性	编号	参考文献	主要准则	通用性	复杂性
1	[Oberti 2001]	D-1	所有算法	中	5	[Prati 2003]	D-1	所有算法	高
2	[Cavallaro 2002]	D-1,D-2	视频 ¹	中/高	6	[Rosin 2003]	D-1	视频 ¹	中
3	[Udupa 2002]	D-1	所有算法	中	7	[Carleer 2004]	D-1, D-3	多目标 ³	低/中
4	[Li 2003]	D-1,D-2	所有算法	高 ²	8	[Erdem 2004]	G-1, G-2	视频 ¹	高

续表

编号	参考文献	主要准则	通用性	复杂性	编号	参考文献	主要准则	通用性	复杂性
9	[Kim 2004]	D-1	视频 ¹	中	22	[Philipp 2008]	G-1, G-3	所有算法	中/高
10	[Ladak 2004]	D-1	所有算法	高 ³	23	[Xu 2008]	D-5, D-6	所有算法	中/高
11	[Lievers 2004]	G-1	阈值化算法	中	24	[Zhang 2008]	G-1, G-2	所有算法	高
12	[Niemeijer 2004]	D-1	所有算法	中	25	[Cárdenes 2009]	D-2, D-7	所有算法	中
13	[Renno 2004]	D-1, D-4	所有算法	中/高	26	[Hao 2009]	G-1	所有算法	中
14	[Udupa 2004]	D-1	所有算法	中	27	[Marçal 2009]	D-2	所有算法	中
15	[Cardoso 2005]	D-2	所有算法	中	28	[Polak 2009]	D-1, D-2, D-5	所有算法	高
16	[Chabrier 2006]	G-1, G-2, D-1	所有算法	中	29	[Qu 2010]	D-5	所有算法	中
17	[Jiang 2006]	D-2	所有算法	中	30	[Casciaro 2012]	D-1	所有算法	高
18	[Ortiz 2006]	D-1	所有算法	高	31	[Khan 2013]	D-8	所有算法	高
19	[Udupa 2006]	D-2	所有算法	中	32	[Peng 2013]	D-5	所有算法	中
20	[Ge 2007]	D-3	所有算法	低	33	[Pont-Tuset 2013]	D-1	所有算法	高
21	[Unnikrishnan 2007]	D-2	所有算法	中					

注 1: 视频指只能用于视频分割, 因为借助了时间信息

注 2: 在评价中需要结合用户观察结果进行判断

注 3: 适合评价图像中有较多目标的分割结果

5.5.3 准则的实验比较

本小节对 5.3 节介绍的一些评价准则进行实验比较。在选取进行比较的准则时有如下考虑。首先各种直接分析方法所用的模型都不相同, 不具有可比性, 所以仅考虑试验评价准则。在 7 组试验评价准则中, 基于形状因子的准则只适合于研究阈值化算法, 不适用于其他类算法。而且它实际上并不能真正描述目标外形[Zhang 1992b], 而只是描述了目标边界的光滑程度; 基于目标计数一致性的准则虽可适用于各种算法, 但并不总是适合于各种实际场合, 这是因为当分割结果比较接近最优时(一般着重研究的也都是这种情况), 实际分割得到的目标个数与图像中原有的目标个数通常是吻合的, 存在的问题只是每个目标并不一定分割得完全正确。在这种情况下, 目标计数一致性已不能用来鉴别分割结果的优劣了。考虑到以上因素, 下面在其余 5 组准则中各取一个准则作为代表进行比较研究。

(1) 准则甲: 区域间对比度。它按式(5.3.3)计算。

(2) 准则乙: 区域内均匀性。它按式(5.3.4)计算。但为了与其他准则相比较只计算该式后面的求和。这样在分割结果较好时所有的指标值都较小, 反之则都较大。

(3) 准则丙: 像素距离误差。它按式(5.3.8)计算。与式(5.3.6)相比, 其好处是避免了主观选择式(5.3.6)中的 p 系数所带来的不确定性。

(4) 准则丁: 像素数量误差。它按式(5.3.9)计算。

(5) 准则戊: 最终测量精度。它按式(5.3.12)计算, 这里选择目标面积作为特征。

对不同准则的比较可借助对一系列分割图的评价来进行。具体就是采用以上各个准则分别评判一系列相同的分割图,算得各指标的一系列值,从这些值的变化中可得到各准则不同的特性,然后通过比较特性来比较准则的优劣。整个实验安排与例 5.3.1 研究目标特征的评价性能有些类似,只是这里借助同一个目标特征,比较了 5 个评价准则。

图 5.5.1 给出了若干实验得到的分割图,其中标 A 的图为原始图像(8 号圆目标),标 B~F 的图为分别用阈值 112、122、124、126、136 分割得到的结果。由图 5.5.1 可以看出,分割结果随阈值的变化呈现由差到好又由好到差的变化,好的评价准则随参数的变化也应有相应的趋势。

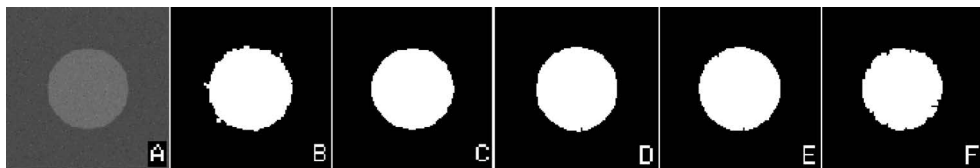


图 5.5.1 用不同阈值分割得到的实验结果

表 5.5.3 给出了一组 5 种评价准则比较的实验结果。其中,各列的标号对应分割图像时所用的阈值,由 5 个指标得到的 5 组值排成 5 行。表 5.5.3 中的数值已分别归一化到 $[0,1]$ 区间以方便比较其相对大小。

表 5.5.3 评价准则比较数据

准则	112	114	116	118	120	122	124	126	128	130	132	134	136	138
甲	0.989	0.994	0.997	0.997	0.998	0.999	0.999	0.999	0.999	1.000	0.999	0.998	0.997	0.995
乙	1.000	0.897	0.858	0.846	0.821	0.808	0.804	0.800	0.800	0.800	0.808	0.825	0.854	0.906
丙	0.705	0.538	0.454	0.415	0.362	0.292	0.260	0.238	0.290	0.382	0.466	0.583	0.719	1.000
丁	0.578	0.340	0.242	0.202	0.154	0.100	0.079	0.066	0.099	0.170	0.254	0.395	0.573	1.000
戊	0.526	0.340	0.241	0.203	0.149	0.092	0.042	0.017	0.077	0.161	0.252	0.395	0.573	1.000

图 5.5.2 以曲线的形式给出了表 5.5.3 中的数值,这样可以比较直观地研究它们的变化趋势和特点。

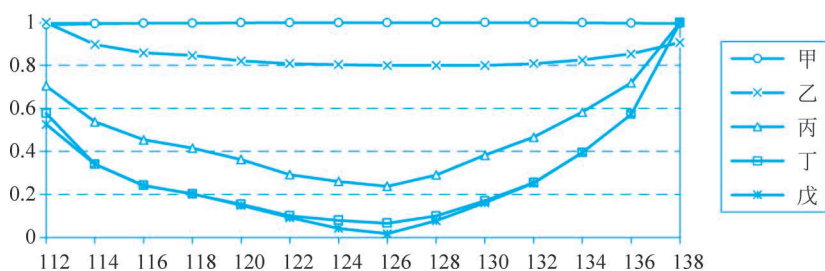


图 5.5.2 评价准则比较曲线

现在来分析表 5.5.3 中的数据和图 5.5.2 中的曲线。可从两方面对图 5.5.2 进行分析。

1. 各曲线的谷的深度

因为最大峰值都已归一化到 1,所以指标的动态范围就由谷的深度值来决定。谷越深表示动态范围越大,而动态范围越大则可表示的不同分割结果越多越细,或者说对分割结果变化的描述比较灵敏。这 5 个准则按动态范围从大到小排列是戊、丁、丙、乙、甲。最终测量精度最好,最差的是区域间对比度,其曲线上的谷值和峰值几乎一样,所以很难据此分辨出不同的分割结果。

2. 各曲线的形状

曲线的形状反映了曲线的走向趋势。在这点上大部分曲线类似,都是先单调下降然后单调上升,这与分割图像排列的情况一致。注意,戊曲线与丁曲线在许多地方都重合,但在接近谷底时(即对应接近最优的分割效果时),戊曲线相比丁曲线下降更多更快(这在表 5.5.3 中也可看出)。这说明利用最终测量精度可以更好地评判接近最优的分割结果,这在实际中是很有意义的。

归纳起来这些准则的性能很不一样,最终测量精度最好,接下来是像素数量误差,而像素距离误差又次之,最后是区域内均匀性和区域间对比度。虽然区域内均匀性和区域间对比度在使用中不需要利用参考图,但它们的评价能力较差。事实上,它们对分割的评价主要还是比较主观的,其他一些实验结果也支持上述结论[罗 1997]、[Xue 1998]。例如,图 5.5.3 给出了一组图像[Xue 1998],图 5.5.3(a)是一幅待分割的原始图,图 5.5.3(b)给出了使用根据区域内均匀性最好而选取的阈值进行分割得到的结果,可见区域内均匀性最好并不代表好的分割结果。图 5.5.3(c)给出了使用根据形状因子最优而选取的阈值进行分割得到的结果,同样分割效果并不理想,可见形状因子的数值并不能正确反映实际分割结果的优劣。

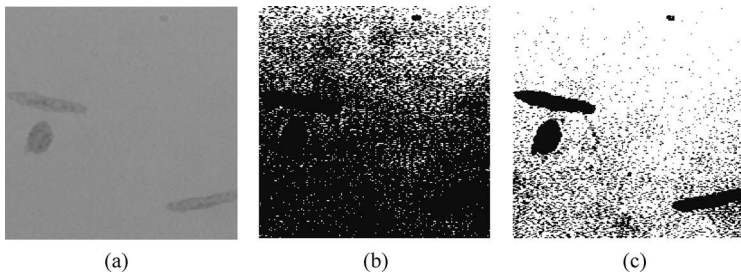


图 5.5.3 区域内均匀性和形状因子的评价能力示例

最后总结一下,从以上实验结果和讨论看,基于差异的准则在精确度和动态范围方面都要比基于优度的准则强,其中最终测量精度准则能提供最大的动态范围,并能最精确地描述分割算法在接近最优分割时的性能。

5.6 分割评价的进展

对分割评价的研究还在不断推进中,下面讨论几方面的工作。

5.6.1 分割算法的评估

在[Wallner 2019]中,对 6 种半自动医学(颅颌面外科)图像分割算法(使用开源代码)在 3 个平台上进行了评估。该研究和过程具有或共享了当前图像分割评价的两个代表性特征。

(1) **监督评价:** 通过对分割算法和真值分割进行比较,评估分割方法的质量和准确性。真值分割的结果通常是在人类专家的帮助下获得的。例如,在该任务中,由临床专家根据观察目标的解剖结构进行判断。

(2) **限于特定的应用领域:** 这与早期的工作不同,早期工作多只考虑分割算法本身,而不考虑要分割的图像和/或目标。在这项工作中,考虑了特定的颅颌面外科图像,并选择下颌骨作为目标(解剖结构)。还可以找到更多类似的例子,如具体到血管[Moccia 2018]、高空间分辨率遥感图像[Chen 2018b]或肾皮质图像中的组织结构[Jayapandian 2021]。

1. 分割过程概述

在该研究中,使用了基于 CT 的颅骨 3-D 模型,并选取了一个没有牙齿的完整下颌骨示意图,由于其在创伤和重建手术中的临床相关性,所以将下颌骨作为解剖结构。

在该研究中,考虑到实验对用户的易用性、免许可证和可重复性,选择了一些开源分割算法。更多的考虑是它们在功能上的稳定性以及与许多软件平台的兼容性。

在该研究中,选择了一些易于下载、安装和使用的平台。更多的考虑是,它们非常广泛地应用于医学成像,而且提供了良好的文档和用户支持。

综合以上考虑,选择了 3 种平台和 6 种分割算法,如表 5.6.1 所示。

表 5.6.1 多平台和多分割算法

平 台	分 割 算 法
3D Slicer	GrowCut, 鲁棒统计分割器(RSS)
MITK	3D 区域生长, Otsu & Picking
MeVisLab	坎尼分割, 大地测量分割器

2. 评价准则和过程

该研究中使用的评价准则是骰子得分系数(DSC)和豪斯道夫距离(HD)。

DSC 描述的是两个分割目标 O_1 和 O_2 之间的(体积)一致性[Sampat 2006]。它测量 O_1 和 O_2 之间的相对体积重叠。如果将 $V(\cdot)$ 定义为目标内体素的体积,则 $V(\cdot)$ 可以通过计算体素的数量,然后将该值乘以体素尺寸来获得。所以有

$$DSC = \frac{2 \cdot V(O_1 \cap O_2)}{V(O_1) + V(O_2)} \quad (5.6.1)$$

两个分割目标体 O_1 和 O_2 之间的豪斯道夫距离(可见附录 A)由目标的边界体素之间的欧氏距离定义[Huttenlocher 1993]。定义 $h(O_1, O_2)$ 为从第一个集合 O_1 中的任何点到第二个集合 O_2 的最小欧氏距离, $h(O_2, O_1)$ 为从第二个集合 O_2 中的任何点到第一个集合 O_1 的最小欧氏距离,这样,两个集合(O_1 和 O_2)之间的 HD(用 $H(O_1, O_2)$ 表示, $H(O_1, O_2) = H(O_2, O_1)$)是这两个距离的最大值:

$$h(O_1, O_2) = \max_{o_1 \in O_1} [d(o_1, O_2)] = \max_{o_1 \in O_1} \left[\min_{o_2 \in O_2} \|o_1 - o_2\| \right] \quad (5.6.2)$$

$$H(O_1, O_2) = \max[h(O_1, O_2), h(O_2, O_1)] \quad (5.6.3)$$

在评价过程中,获得了两个专家的帮助以确定分割真值。从专家 A 获得 10 个真值分割集,从专家 B 获得 10 个真值分割集,以及从 6 个算法中的每一个都获得 10 个算法分割集。将这 80 个分割集在 3 个评估中进行比较:

- (1) 算法分割与真值分割 A;
- (2) 算法分割与真值分割 B;
- (3) 真值分割 A 与真值分割 B(这用于避免导致真值分割变化的偏差)。

除了 DSC 和 HD 之外,真值分割 A 和真值分割 B 之间的一致性也由体积和体素值确定。对上述 3 种评价案例进行描述性统计计算,以总结测量和评估参数(最小值、最大值、平均值和标准差)。

此外,借助分析统计方法以相互比较真值分割(A,B)。这些分析统计方法包括计算成对 t 检验(p)和皮尔逊积矩相关系数(r)[Rogers 1988]。在真值分割(A,B)之间计算概率值。 p 值低于 0.05($p < 0.05$)被认为是显著的。在这项研究中,实验表明,当比较真值分割时,体积和体素值的皮尔逊积矩相关系数(r)接近值 1($r > 0.99$)。

通过该研究给出的一般结论包括:

- (1) 由于不存在通用的分割理论,因此分割方法/算法的评价应考虑许多因素,包括应用领域;

(2) 由于已经提出了许多具有各种原理的分割方法/算法,所以看起来似乎任何评价研究都不能完全涵盖这些方法/算法的所有方面;

(3) 由于分割方法/算法经常利用应用领域中的某些先验知识来提高分割性能,因此限制在指定的应用领域将使评价的方法/算法更具有可比性。

5.6.2 像素分类的评价准则

从对血管图像的分割评价入手,[Moccia 2018]对其中的方法、数据集和评价准则进行了调查和评述,其中讨论了图像分割评价的一些准则。其中有些思想和方法可以扩展到更广义的图像分割问题。以下仅讨论评价准则部分。

图像分割经常被认为是像素分类任务。在许多医学图像应用中,通常根据黄金标准(GS)或真值(GT)来评价分割性能。这需要由一个或几个领域专家执行手动分割,并需要根据不同的策略来组合多个分割。一种允许合并被分割结构的先验信息或估计不完美或有限参考标准的方法称为同时真实性和性能水平估计(STAPLE)[Warfield 2004]。

从像素分类的角度来看,可以计算 4 个分割结果(4 个指标),其中真和假(正和负)分别指属于与 GT 分割一致的目标和背景的像素。

(1) 真阳性(TP): 分割目标像素(s_o)被分类为 GT 目标像素(g_o)类。

(2) 真阴性(TN): 分割背景像素(s_b)被分类为 GT 背景像素(g_b)类。

(3) 假阴性(FN): 分割背景像素(s_b)被分类为 GT 目标像素(g_o)类。

(4) 假阳性(FP): 分割目标像素(s_o)被分类为 GT 背景像素(g_b)类。

借助图 5.6.1 可以说明它们之间的关系。

1. 基本评价准则

基本评价准则是上述 4 个指标的简单组合。最常用的评估标准是准确性(Acc,也称为查准率)、敏感性(Se,也称为查全率)和特异性(Sp)。通常,Se 越高,Sp 越低,FP 越高,则分割结果越好[Powers 2011]。此外,常用的其他评价准则还包括: FP 率,等于 $1 - Sp$; 正预测值(PPV),也称为精度,即 TP 在 $TP + FP$ 中的比例; 以及负预测值(NPV),即 TN 在 $TN + FN$ 中的比例。表 5.6.2 列出了这些基本评价准则。

	g_o	g_b
s_o	TP	FP
s_b	FN	TN

图 5.6.1 TP、FP、FN、TN 的关系

表 5.6.2 基本评价准则

准 则	公 式	描 述
Acc	$(TP + TN) / n$	真实结果(TP 和 TN)占评价总数(n)的比例
Se	$TP / (TP + FN)$	正确识别的阳性(TP 和 FN)比例
Sp	$TN / (TN + FP)$	正确识别的阴性(TN 和 FP)比例
FP 率	$FP / (TN + FP)$	错误识别的阴性(TN 和 FP)比例
PPV	$TP / (TP + FP)$	PPV 给出了像素属于血管的可能性的估计,假设算法将其分类为正
NPV	$TN / (TN + FN)$	NPV 对应于像素不属于血管的可能性,假设算法将其分类为负值

2. 扩展评价准则

扩展评价准则是上述 4 个指标的衍生组合。最常用的评价准则是接收机操作特性(ROC)曲线(精确召回曲线或 PPV-Se 曲线)下的面积,并表示为 AUROC; 马修斯相关系数(MCC); 科恩系数 κ ; 骰子得分系数(DSC),也称为 F_1 分数; 以及豪斯道夫距离(HD)。马修斯相关系数一般情况下是皮尔逊相关系数[Rodgers 1988]的二分形式。科恩系数 κ 是衡量评估者之间一致性的一个指标。表 5.6.3 列出了这些扩展评估标准。

表 5.6.3 扩展评价准则

准 则	公 式	描 述
AUROC	$\text{AUROC} = \sum_{p \in [0,1]} r(p)$	指示分类器将随机选择的阳性实例的排名高于随机选择的阴性实例的概率
MCC	$\text{MCC} = \frac{(\text{TP} \cdot \text{TN}) - (\text{FP} \cdot \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$	相关系数的一种变型
κ	$\kappa = \frac{\text{Acc} - p_e}{1 - p_e}$	其中, p_e 是假设的机会一致概率, 等于 GS 生成阳性的概率乘以分割算法生成阳性的可能性
DSC	$\text{DSC} = \frac{2\text{card}(O \cap G)}{\text{card}(O) + \text{card}(G)} = \frac{2\text{TP}}{\text{FP} + \text{FN} + 2\text{TP}}$	两个聚类 O 和 G 的交集元素的数量与平均标签图像的比率, 其中 O 是分割目标, G 是相应的 GS 目标
HD	$\text{HD} = \max[\sup_{o \in O} \inf_{g \in G} d(o, g), \sup_{g \in G} \inf_{o \in O} d(o, g)]$	$\sup$ 表示上确界, $\inf$ 表示下确界, d 可以是任何距离度量

3. 更多评价准则

所有上述准则都基于分割图像(S)和 GS(G)的目标之间的像素比较。事实上, 图像分割不仅用于对图像的划分, 还用于提取感兴趣的连通组元(目标)。所以, 对图像分割进行基于像素的评价, 也需要从两方面着手。

一方面, 应该计算 GS 目标和分割目标之间的点对点对应关系。例如, 在血管分割的特定情况下, 可以使用血管中心线上点的对应关系。由此可得出 3 种不同的中心线重叠度量作为评价准则。

(1) 重叠(OV)测量: 跟踪观察者标注的整个血管的能力。

(2) 直到第一个误差的重叠(OF)测量: 从 GS 中心线的第一个点到分割中心线上对应点的最小半径。

(3) 与血管的临床相关部分的重叠(OT)测量: 仅测量对直径大于或等于 1.5mm 的血管进行分割的能力。也就是说, 确定最接近参考 GS 末端且半径大于或等于 0.75mm 的点以限制相关部分。

另一方面, 还可以利用目标的各种特性[Zhang 1996a], 例如, 连通性(C)、面积(A)和周长(P)。因子 C 通过比较 G 中连接组元的数量和 G 中目标像素的总数来惩罚破裂的分割[Gegúndez-Arias 2012]。因子 A 测量分割目标(O)和 GS(G)中目标之间的重叠程度。在 A 的计算中, 执行了一些形态学膨胀操作, 这使得 A 与 DSC 相比, 对实际分割和 GS 之间的细微差异不太敏感。因子 P 可以测量分割血管长度和 GS 血管长度之间的一致性。

表 5.6.4 列出了上述两组评价准则。

表 5.6.4 基于像素之间比较或点对点对应的评价准则

准 则	公 式	描 述
OV	$\text{OV} = \frac{\text{TPR}_{\text{ov}} + \text{TPM}_{\text{ov}}}{\text{TPR}_{\text{ov}} + \text{FN}_{\text{ov}} + \text{TPM}_{\text{ov}} + \text{FP}_{\text{ov}}}$	TPR_{ov} 指 GS 中心线上的点, 其与分割中心线上对应点的距离小于局部血管半径。该距离大于半径的点标记为 FN_{ov} 。如果 GS 上至少有一个点的该距离小于半径, 则分割中心线上的点标记为 TPM_{ov} , 否则标记为 FP_{ov}
OF	$\text{OF} = \frac{\text{TPR}_{\text{of}}}{\text{NR}}$	它是第一个误差之前的 TPR_{ov} 的数量(TPR_{of})与参考点的总数量(NR)的比率

续表

准 则	公 式	描 述
OT	$OT = \frac{TPM_{ot} + TPR_{ot}}{TPM_{ot} + TPR_{ot} + FN_{ot} + FP_{ot}}$	TPM_{ot} 、 TPR_{ot} 、 FN_{ot} 、 FP_{ot} 按 TPM_{ov} 、 TPR_{ov} 、 FN_{ov} 、 FP_{ov} 计算, 但仅考虑 p_{end} 和 GS 中心线起点之间的点
C	$C = 1 - \min \left[1, \frac{ \text{card}(G) - \text{card}(O) }{\text{card}(G)} \right]$	它评估 G 和 O 之间的碎裂程度。在血管分割的情况下, 预计只有几个相连的组元(理想情况下是一个)
A	$A = \frac{\text{card}\{[d_A(O) \cap G] \cup [d_A(G) \cap O]\}}{\text{card}(O \cup G)}$	d_A 是计算 A 中用于形态膨胀的掩模半径
P	$P = \frac{\text{card}\{[d_P(O) \cap G] \cup [d_P(G) \cap O]\}}{\text{card}(O \cup G)}$	d_P 是计算 P 时用于形态扩张的掩模半径, S 表示同伦骨架化

5.6.3 全景分割的评价

在 2.7 节中介绍的全景分割(PS)分别考虑了物质和事物的分割[Kirillov 2019], 因此物质和事物统一度量将鼓励研究统一任务, 或者换句话说, 评价涉及物质类和事物类的联合任务。

1. 全景质量

全景质量(PQ)是一种适合 PS 的度量(评价准则)。其中考虑了 3 个因素[Kirillov 2019]。

(1) 完整性。它应该以统一的方式处理物质和事物类, 并考虑到联合分割任务的所有方面。

(2) 可解释性。它应该具有便于沟通和理解的可识别的含义。

(3) 简单性。它应该易于定义和实现, 从而可以快速高效地计算。

使用 PQ 可测量全景分割结果相对于真值的质量。它包括两个步骤。

(1) 在分割结果和真值中匹配相应的图像块, 其中图像块是目标的一组像素(在极端情况下, 也可以是一个像素)。

(2) PQ 值的计算(在确定分割结果中的图像块与真值中的图像块之间的匹配之后)。

2. 图像块匹配

当两个匹配的图像块相交时, 它们有可能会匹配。在定义 PQ 时, 要求它们在并集上的交集(IoU)严格大于 0.5。已经证明[Kirillov 2019], 这一要求与全景分割的非重叠属性一起, 可以保证匹配的唯一性。换句话说, 最多可以有一个分割图像块与每个真值的图像块相匹配。

匹配对的 IoU 必须大于 0.5 的要求也符合上述考虑。第一, 这是简单而有效的, 因为两个图像块之间的对应是唯一的并且容易获得。第二, 它是明确和可解释的。注意, 由于唯一性, 对于 $\text{IoU} > 0.5$, 任何合理的匹配策略(包括贪婪和最优)都将产生相同的匹配。实验表明, 具有 $\text{IoU} \leq 0.5$ 的匹配在实践中是罕见的。

3. 计算评价

全景分割包括物质类和事物类的分割。全景分割的评价指标也应该考虑这两类。所提出的 PQ 计算独立地处理每个类, 最终结果是类的平均值。

对于每个类, 根据上述关于匹配唯一性的讨论, 分割图像块和真值图像块可分为 3 组:

- (1) TP(真阳性)——满足匹配的图像块(分割的, 真值的);
- (2) FP(假阳性)——不匹配的分割图像块;
- (3) FN(假阴性)——不匹配的真值图像块。

根据这 3 个集合, PQ 可以定义为(g 是真值图像块, s 是分割图像块)

$$PQ = \frac{\sum_{(s,g) \in TP} IoU(s,g)}{|TP| + |FP|/2 + |FN|/2} \quad (5.6.4)$$

式(5.6.4) 右侧可以分解为两个因子, 即

$$PQ = \frac{\sum_{(s,g) \in TP} IoU(s,g)}{|TP|} \times \frac{|TP|}{|TP| + |FP|/2 + |FN|/2} \quad (5.6.5)$$

第一项可以被称作**分割质量(SQ)**, 而第二项可以被称作**识别质量(RQ)**。第一项是匹配图像块的平均 IoU, 第二项是检测质量的估计, 即 F_1 分数。

值得注意的是, 对 PQ 的分解, 即 $PQ = SQ \times RQ$ 为分析提供了一些信息, 但这两个因子并不独立, 因为 SQ 仅在匹配的图像块上测量。

最后要指出的是, 无论图像块的尺寸如何, 所有图像块都会得到相同的权重。所以, 这样的标准似乎更适合于评价相似尺寸实例的分割。

5.7 基于评价的算法优选系统

分割评价的目的是指导对算法的设计和选择, 改进和提高算法的使用性能。将评价和分割应用联系起来的一种方法是结合人工智能技术, 建立**分割专家系统**以有效地利用评价结果进行归纳推理, 从而把对图像的分割由比较盲目地试验改进层次上升到系统地选择实现层次。

5.7.1 算法优选思想和策略

在本章, 开头曾指出, 尽管已有上千上万种图像分割算法, 但并没有一种适用于所有图像分割任务。前几节关于图像分割评价的讨论更表明, 即使都属于或符合同一模型的图像, 当目标的尺寸、形状等发生变化以及当各种干扰因素改变时, 都会对分割算法提出不同的要求。如何能动态地适应这些变化而且系统地选择恰当的算法来分割图像呢?

前面曾介绍了根据对分割图中目标特征测量的精确度来判断分割算法性能的评价方法和准则。通过这样的评价可以得到待分割图像的特性参数与所用分割算法性能之间相关的知识。进一步借助这种知识的指导, 根据对待分割图像特性的分析和估计, 就可以预测不同算法的分割效果, 从而选择最优算法。这是一种基于分割评价对算法进行优选的基本思路[章 1998b]。

为实现上述算法优选方案而设计的图像分割算法优选系统的框图如图 5.7.1 所示[章 1998b]。系统(图 5.7.1 中虚线框内)的输入包括待分割图(在图像库中)和各种关于分割的知识(在知识库中), 还可包括一系列分割算法(在算法库中)。系统的输出是系统选出的最优算法及由此得到的输出图。图 5.7.1 中点画线框内基本上是一个分割评价系统。通过对输入图进行分割, 对分割结果进行判断, 就能建立图像特性与算法的联系。实际中, 图像特性主要由图像信噪比、图像模糊度、目标与背景的对比度、目标面积、形状和轮廓粗糙度等所决定[Zhang 1992c]。这些特性有些可从原始图像中估计出来, 例如, 对噪声的估计可参见[Olsen 1993], 对模糊图像中边缘参数的估计可参见[Kayargadde 1994]。需要注意的是, 同一个具体应用领域中的各个待分割图像是具有内在联系的, 而且许多图像和目标的特性常可粗略地借助某种图像模型描述。据此也可以根据模型构造图像进行分割以通过评价来获取相应的先验评价知识。

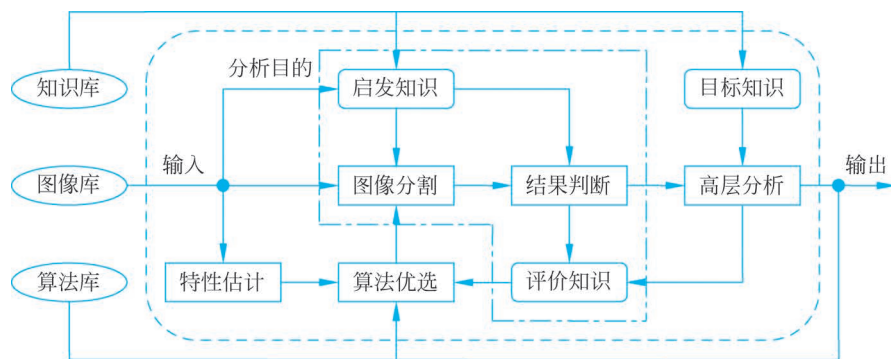


图 5.7.1 图像分割算法优选系统框图

一旦获得了足够的评价知识,对给定的输入图,通过特性估计,就可借助图像特性与算法性能的对应关系进行算法优选[罗 1998]。这个优选过程是一个知识驱动的“假设-检验”反馈过程。首先,根据先验的图像特性估计或测量做出可以产生先验最佳算法的假设;接着,根据使用所选出算法进行分割所得到的分割结果又可以计算后验的图像特性估计,并获得对应的后验最佳算法估计。若先验的假设是正确的,那么后验估计就应该与先验假设相吻合或一致;否则可用后验估计更新先验估计,以激发新一轮的“假设-检验”过程,直到两者满足一定的一致性条件为止。可以看出,这种反馈过程是一种对信息逐渐提取与逼近的过程。算法选择在这个过程中逐步得到优化,最终趋向于最优,并将最优的算法选出来。由于这种反馈方式主要由数据驱动来对分割环节进行自身调整,所以是一种自底向上的处理过程,因而比较迅速和方便。

在这种“假设-检验”方式下,对于分割算法的选择主要是用评价知识导引的。尽管这种指导能力和作用还要在反馈中进行调整,但它与纯粹的“尝试-反馈”模式不同,它并不需要尝试所有候选算法以得到最优解。这里由于图像分割算法的复杂性,算法评价知识的获取常受到各种限制,为此有时需要借助一定的启发性知识,以把主观上对分割性能的理解规则加入分割和评价过程。最后从加强高层知识对分割指导作用的角度考虑,还可利用高层分析手段。这里借助对分割之后目标的认识,通过对与目标区域属性有关特性的分析,由上而下地进一步保证对算法的优选(见图 5.7.1)。

5.7.2 优选系统的实现和效果

分割算法优选系统是一个知识驱动的系统。为提高系统的性能和增进系统的效率,在设计中还引进了启发性知识和基于高层目标分析的反馈知识以作为评价知识的补充,所以在系统中包含多种知识源。上述优选系统也可看作一个控制系统,系统中包含多种控制和反馈机制以管理多种知识源。基于这种考虑,使用了基于公共数据黑板的控制系统结构,将控制知识划分为知识源的形式,各知识源通过公共数据黑板交换信息和协调运行。这里所有知识可归纳成两类:

- (1) 静态知识——包括算法处理方式以及各算子之间的组合关系,采用“框架”结构来表示;
- (2) 动态知识——指系统中与动态系统状态相关的控制知识,用产生式规则及相应的附加过程来表示[Luo 1999]。

在该系统中,公共数据黑板被分为 4 个区[罗 1998]: 图像数据区(放置由分割算法所处理的 2-D 图像数据)、图像特征区(放置特性估计所得到的参数)、控制数据区(放置记录系统优化运行的历史数据)和分割目标区(记录用户指定的分割区域所具有的特征)。

该系统的算法库中已有两种基本的基于边缘检测的分割算法(它们可用于分割灰度梯度

较大的图像)和6种典型的阈值选取算法。这6种阈值选取算法为(参见[罗 1997])改进直方图法、二维最大熵方法、直方图分析法、最小误差法、矩保持法和简单统计法。

通过用图 5.7.1 的评价框架可对上述这些算法进行试验评价,然后可将所得到的评价知识纳入到算法优选系统中。

使用该系统已对几十幅不同类型的图像进行了算法优选试验。在试验中,为便于比较,各算法在取阈值分割前均未采用预处理手段,分割后的后处理方法(利用了形态学开启和闭合,见第 14 章)对各算法都相同。其中一个典型的试验[Luo 1999]使用了尺寸为 256×256 px,灰度为 256 级的癌细胞切片图(见图 5.7.2(a),图 5.7.2(b)为图 5.7.2(a)的灰度统计直方图)。

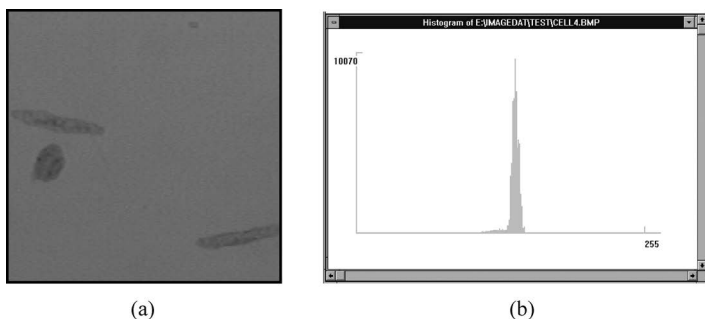


图 5.7.2 癌细胞切片测试图及其直方图

图 5.7.3(a)~图 5.7.3(f)依次是为了验证系统优选效果用上述 6 种阈值分割算法(按前面介绍的次序排列)对图 5.7.2(a)分割而得到的结果。比较图 5.7.3 中的各图,可知对应图 5.7.3(c)的直方图分析法给出的分割结果最好,其次是对应图 5.7.3(e)的矩保持法给出的分割结果。试验中系统经过两次尝试选取了直方图分析法作为输出算法。

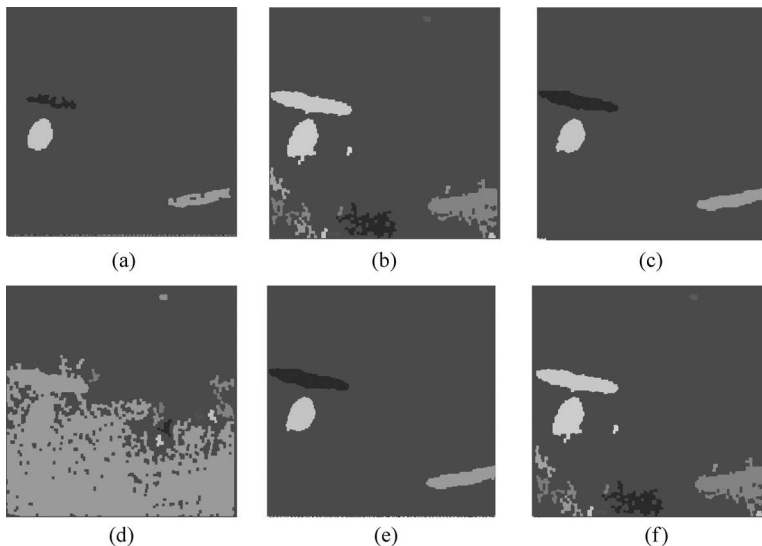


图 5.7.3 用 6 种阈值分割算法分别分割图 5.7.2(a)得到的结果

最后需要指出的是,在对其他一些图像的试验中,系统一般经过两到三次尝试都能自动选出适合于给定图像的最优分割算法或效果与最优算法最接近的亚最优算法。这表明系统在设计预期的范围内能达到比较满意的优选效果,与穷举方法相比效率有很大提高,而且可自动完成。另外试验结果表明,6 种算法都曾入选最优算法,这也验证了优选算法的必要性。其他类算法也可以结合到该系统中。



彩图

总结和复习



随堂测试

