第 27 卷第 1 期　　南 京 气 象 学 院 学 报　　Vol. 27 No. 1
2004 年 2 月　　Journal of Nanjing Institute of Meteorology　　Feb. 2004

文章编号: 1000-2022(2004) 01-0050-05

一种基于统计分布的图像分割方法

罗维佳,　谢顺平,　都金康
(南京大学 城市与资源学系, 江苏 南京　210093)

摘　要: 为了使数字高程模型(DEM)图像显示时各种颜色所占面积大体相等, 提出了一种图像分割多阈值确定方法。该方法先用较多的组数对数据点按高程分组统计, 在此基础上合并数据点较少的相邻组, 得到分布较为均匀、组数较少的二次分组, 提取二次分组的组限作为阈值。试验和分析证明, 用该方法获取的阈值能较好地分割图像, 节省时间和空间。

关键词: 图像分割; 阈值; 统计; 数字高程模型
中图分类号: TN911.73　　文献标识码: A

地理信息系统软件在二维彩色显示数字高程模型(DEM)数据时, 通常提供两种分割区域方式: 一是手工输入高程分割阈值, 一是列出一些能确定高程阈值的分布曲线。有时希望分割后的各区域面积大体相等, 使得显示的图像上各种颜色所占的面积大体相等, 达到较好的视觉效果, 按上述两种方式分割常常得不到满意的效果。在显示时, 按照最大高程值与最小高程值之差进行平均分割, 这是一种获取阈值的方法, 通常效果并不理想, 因为各地的 DEM 高程统计分布并不相同, 很少出现直线分布的情形。同样的原因, 直接通过各种曲线来确定的阈值也不理想, 除非曲线是通过 DEM 数据拟合而得到的[1], 然而这步工作不是必要的。一种可行的方法是先对 DEM 数据点按高程值进行排序, 对排好序的数据点按点数平均分割而获取阈值, 这种方法存在的问题是运算耗时, 需要大量的内存空间。为了实现对 DEM 的多阈值等面积分割, 本文提出一种视觉效果较为理想的阈值确定方法, 它针对特定区域 DEM 的实际高程的统计分布来确定多个分割阈值, 不需拟合曲线, 只占用微小的运算时间和内存空间。在实现过程中有两个关键问题需要考虑: (1) DEM 的高程取值范围因地而异, 取值精度随任务不同而变化, 分多少组对高程值进行统计比较合适? (2) 由于 DEM 区域分割的数目为 $m > 2$, 如果每次都累加到大于或等于图像面积的 m 分之一, 则由于先分的区域可能超过各自的应有的 m 分之一, 使得最后几个区域的面积为 0, 实际得到的区域数目小于 m, 未能达到分割目的。本文用算法解决了上述第 2 个问题, 在试验中对第 1 个问题进行了讨论。

收稿日期: 2003-03-10; 改回日期: 2003-05-20
基金项目: 国家自然科学基金资助项目(40171015)
作者简介: 罗维佳(1971-), 男, 云南澄江人, 硕士生.

1　基本思想及算法

基本思想是先对 DEM 数据按高程进行间距较小、组数较多的分组统计, 根据统计结果把次数较少的相邻组合并, 得到组数较少的分组, 并使得各组的次数大致相等。由于各组的次数代表了各组对应的高程范围(该组的上下组限) 内的点数, 而点数的多少对应着面积的大小, 当各组的次数大致相等时, 其所对应的面积也大体相等, 因此可以用第 2 次分组的组限作为阈值。

本算法的实现步骤如图 1 所示。读取DEM 数据及相关特征值后, 即可计算出小组的组距 d_c, 根据组距 d_c、最小高程值 Z_{min}和第 i 点高程值 data[i] 确定 i 点应归入的小组 k, 从而实现了第 1 次分组。第 2 次分组时, 依次累加相邻小组的次数, 直到累加值等于或超过平均数, 把这些相邻小组合并成一个颜色组, 提取最后一个小组的上限作为一个阈值; 然后重新累加剩余的相邻小组, 直到累加值等于或超过剩余部分的平均数, 又得到一个颜色组及相应阈值; 如此下去, 便可确定 $m-1$ 个阈值, 把有意义的点分为点数大体相等的 m 组。

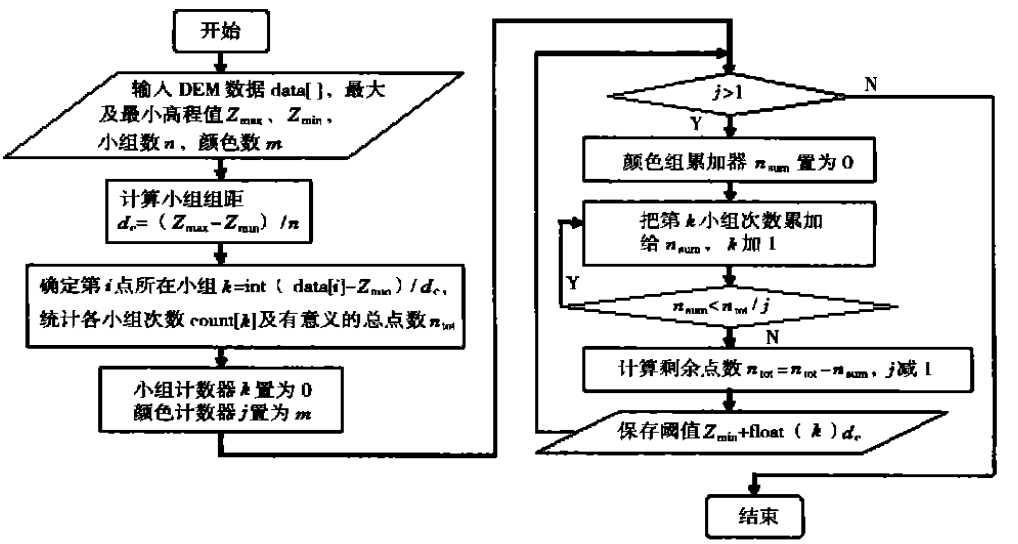

图 1　图像分割算法流程

Fig. 1　Flow chart for image segmentation algorithm

通常情况下, n 的值比 m 的值大得多, 即第 1 次分组的组数比第 2 次分组的组数(欲使用的颜色数) 大得多。这样做的目的是降低第 1 次分组时次数较高组的次数, 以使第 2 次分组时各组的次数差距不要太大。这样就保证了即便出现偏态分布(在某一个高程范围内集中了大量的点), 也能得到次数较为平均的分组。这种分组没有必要实现绝对的平均, 因此一方面 n 的值没有必要很大, 远远小于总点数, 另一方面在确定阈值时, 如果每次与累加值 n_{sum} 进行比较的值都为 n_{tot}/m(这时 n_{tot} 为固定不变的总点数), 就会出现上述第 2 个问题。为了避免出现这种情况, 采用剩余点数的平均数 $n_{tot}/(m+1-j)$ (这时 n_{tot} 为经过前面分组后所剩余的点数) 与 n_{sum} 进行比较。

2　试验及结果分析

本研究在微机上用 Visual C+ + 6.0 结合 OpenGL 图形库函数编程试验[2], 数据采用了

南京附近的江宁地区和浙江涌江流域的黄土岭的格网 DEM 数据, 这里以江宁地区数据进行说明。江宁地区的 DEM 格网间隔为 100 m, 总的数据规模为 574×540= 309 960 点, 其中包含不少无意义的点, 它们的高程值为 0 m, 剔除这部分点外, 剩余点数为 205 384。在有意义的点中, 最小高程值为 10.0 m, 最大高程值为 414.0 m, 所有点的高程值精确到 1 m。

对于有意义的点, 欲通过红、品红、黄、蓝、青、绿 6 种颜色来显示, 即欲把有意义的点分为点数大体相等的 6 组, m= 6, 各颜色组点数 x 的平均数为 $\bar{x}$= 205 384/ 6= 34 231, n 表示第 1 次分组时的小组数。使用不同的 n, 按上述步骤进行处理, 得到表 1 结果, 其中标准差计算公式[3] 为: = $\sqrt{\sum (x - \bar{x})^2 / m}$。

表 1 小组数 n 取不同值时各颜色组对应的次数及标准差

Table 1 Frequeces and standard deviations of color groups at different n

n	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	第 6 组	标准差
6	186 114	14 495	3 859	786	112	18	68 110
60	81 175	36 539	23 734	29 717	17 978	16 241	22 089
120	74 645	38 979	27 824	21 420	21 836	20 680	19 141
180	72 369	42 588	26 491	22 011	21 943	19 982	18 653
240	70 052	43 572	27 824	21 420	21 445	21 071	17 852
404	67 684	45 940	27 824	21 420	21 445	21 071	17 311
500	67 684	45 940	27 824	21 420	21 445	21 071	17 311

从表 1 可看出, 当 n= 6 时(m= n), 绝大部分点集中在第 1 组内, 其余各组分布较少, 分布极为不均匀, 各组次数与平均数 34 231 差距较大。从图 2 可看到, 对应于第 1 组的绿色占据了绝大部分面积, 视觉效果差。这实际相当于直接用最大高程值与最小高程值之差进行 6 等分的平均分割。

当小组分组数 n 扩大到 60 后, 由于每小组的平均次数下降, 根据小组合并所得各颜色组的次数较为均匀, 各组次数与平均数 34 231 的差距大大缩小。从图 3 可看到, 对应于第 1 组的绿色所占面积大幅缩小, 对应于其他组的颜色所占面积增加明显, 视觉效果得到显著提高。

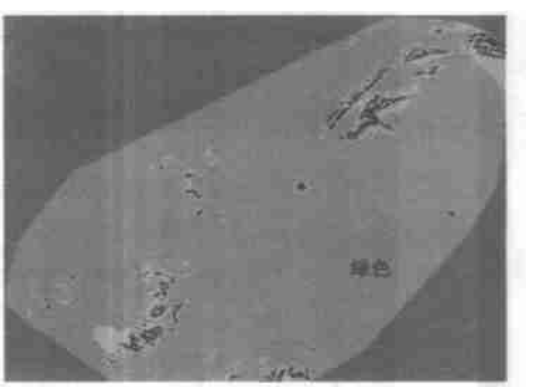

图 2 n= 6、m= 6 时所得图像

Fig. 2 Resultant image when n= 6 and m= 6

从图 4 可看到, 当 n 进一步扩大到 404 时, 绿色所占面积继续缩小, 但所占面积比重仍然超过其他颜色, 总体来看各种颜色所占面积更为均匀, 视觉效果更好。由表 1 可看出, 随着小组数的增加, 占有绝对比重的第 1 组数目呈下降趋势, 其余各组呈增加趋势, 6 个颜色组都向平均数 34 231 接近, 表明离散程度的标准差也越来越小, 说明各组次数越来越集中在平均数附近, 这些正是我们所希望的。

从表 1 还可看出两点: (1) 当小组数为 404 和 500 时, 所得各颜色组的次数及标准差都是一样的。本次试验使用的数据中, 最小高程值为 10.0 m, 最大高程值为 414.0 m, 又由于数值

精度为 1 m, 因此可理解为高程值都是 10 到 414 之间的整数。当小组数为 404= 414- 10 时, 已达到能细分的极限, n 取大于 404 的值已无意义。若以　表示数值精度, 可得有意义的 $n \leq (Z_{max} - Z_{min})/$　。(2) 当小组数以 60 的倍数递增时, 标准差虽然在减少, 但并没有以某个值的倍数递减, 出现的情况是一开始下降很快, 后来逐渐减缓。这也表明当 n 达到一定大小后, 欲通过扩大 n 的值来增加均匀程度已意义不大。

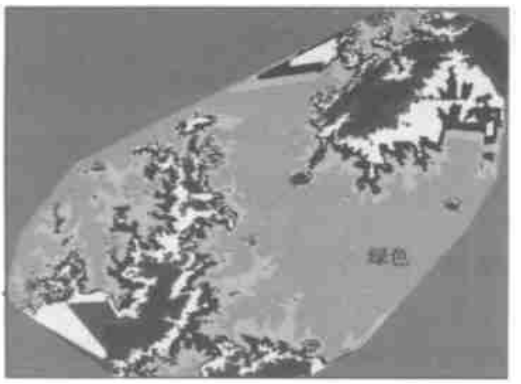

图 3　n= 60、m= 6 时所得图像

Fig. 3　Resultant image when n= 60 and m= 6

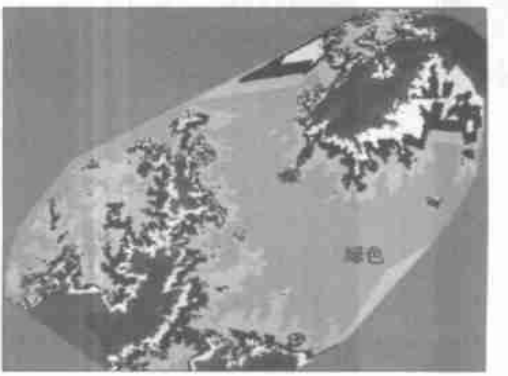

图 4　n= 404、m= 6 时所得图像

Fig. 4　Resultant image when n= 404 and m= 6

3　算法性能分析

由上述算法步骤可看出, 由于 DEM 数据量很大, 该算法主要的时间用于实现第 2 步: 遍历每一个 DEM 高程点, 根据其高程值把它划入相应的小组内。若以对每一个高程点的统计为基本操作, 整个 DEM 共有 n 个点, 则该算法的时间复杂度为 $O(n)$。这远远优于排序方法, 因为即便平均计算时间最短的快速排序的时间复杂度也为 $O(n\log_2 n)$[4]。

格网 DEM 数据的一个特点是: 每一个点的平面位置隐含在该点所对应的行列号中[5]。如果采用排序方法提取阈值, 为了保留每个点的原始顺序, 必须要占用一块与 DEM 数据量相当的内存空间, 用来复制 DEM 数据或建立索引, 以供排序使用, 因而其空间复杂度为 $O(n)$。采用本文算法实现时, 不需要复制 DEM 数据或建立索引, 主要的空间占用是统计各小组次数的数组 count[i], 通常情况下该数组元素个数为 500 时就能得到较理想的分割效果(数值精度较高的黄土岭地区的数据分割结果十分接近绝对平均)。即便该数组大到包含 1 000 个元素, 相对于 DEM 的数据量来说仍然是很小的, 因此其空间复杂度可认为是 $O(C)$, C 表示一较小常量。

4　结　论

(1) 采用本文所提出的算法获取的图像分割阈值, 能大体平均地分割图像, 获得较好的视觉效果。

(2) 本文所提出的算法在性能上远远优于排序方法, 运算速度快, 占用内存空间小。

(3) 随着细分小组数 n 的增加, 二次分组的结果越来越均匀。有意义的细分小组数 n 的取值范围是: $m \leq n \leq (Z_{max} - Z_{min})/$　, 当 $n > (Z_{max} - Z_{min})/$　时, 最后所得结果与 n= (Z_{max}- Z_{min})/ 时相同, 通常 n= 500 时已能得到较好的结果。

(4) 在对灰度图像(如遥感影像) 进行伪彩色增强时, 如果欲把图像覆盖区域划分为面积大体相等的几类, 也可按本文方法确定灰度阈值。

参考文献:

[1] Castleman Kenneth R. 数字图像处理[M]. 北京: 电子工业出版社, 2002.
[2] Wright Richard S, Sweet Michael. OpenGL 超级宝典(第2版)[M]. 北京: 人民邮电出版社, 2001.
[3] 李洁明, 祁新娥. 统计学原理[M]. 上海: 复旦大学出版社, 1995.
[4] 殷人昆, 陶永雷, 谢若阳, 等. 数据结构[M]. 北京: 清华大学出版社, 1999.
[5] 黄杏元, 马劲松, 汤 勤. 地理信息系统概论(修订版)[M]. 北京: 高等教育出版社, 2001.

An Image Segmentation Algorithm Based on Statistic Distribution

LUO Wei-jia, XIE Shun-ping, DU Jin-kang
(Department of Urban and Resources Sciences, Nanjing University, Nanjing 210093, China)

Abstract: An image segmentation algorithm is proposed in this paper to display digital elevation model(DEM) image with colors which occupy respectively an approximately equal area in display window. Firstly, the data set is classified into many groups according to elevation values; secondly, the groups adjacent with few data points are combined, when each group has approximately equal number of data points, and their limits are taken as the thresholds. The results of experiment and analysis show that the algorithm is efficient as well as time- and space-saving.

Key words: image segmentation; thresholding; statistics; digital elevation model