

# 基于谱聚类的两阶段颜色量化算法

谷瑞军 叶 宾 须文波

(江南大学信息工程学院, 无锡 214122)

**摘 要** 颜色量化是进行图像处理和图像分析的重要技术之一, 可以被广泛地应用到图像分割、图像压缩和图像识别中。首先利用高效的二分  $K$  均值聚类进行粗略量化, 然后使用基于加权距离的谱聚类进行再次量化。实验结果表明, 和其他常见量化算法相比, 两者的结合使得新方法在运算速度和量化质量上都取得了不错的结果, 而加权距离的引入, 有效地解决了传统算法将包含像素个数少但重要的颜色进行错划分的问题。

**关键词** 颜色量化 谱聚类 二分  $K$  均值

中图分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(2007)10-1922-04

## A Two-phase Color Quantization Approach Based on Spectral Clustering

GU Rui-jun, YE Bin, XU Wen-bo

(School of Information Technology, Jiangnan University, Wuxi 214122)

**Abstract** Color quantization or color reduction is an important technique for image analysis and has been widely used in image segmentation, image compression and image recognition. Firstly, on bisecting  $K$ -means is used to quantize image roughly and then we refine the image by improved spectral clustering based weighted distance. The stability and quickness of bisecting  $K$ -means and adjustable weight make our approach an attractive one. Experimental results show that our approach performs better than octree algorithm in quantized quality and has a less computation complexity than  $K$ -means algorithm. For special image, which includes one important color but with only a few pixels, traditional approaches usually lose the important color, but our approach can deal with it by introducing the weight for distances between pixels.

**Keywords** color quantization, spectral clustering, bisecting  $K$ -means

## 1 引 言

颜色量化是进行多媒体处理的重要技术之一, 已被广泛应用到图像分割<sup>[1]</sup>、图像压缩<sup>[2]</sup>和视频压缩等领域。具体是指用  $c$  种颜色重建包含  $N(N \gg c)$  种颜色的图像, 并使重建误差最小。颜色量化的过程有两步: 第1步是选择合适的调色板; 第2步是用调色板中的颜色代替像素的原来颜色重建图像。常见的量化算法主要有两大类<sup>[3]</sup>: 分割算法和聚类算法。典型的分割算法有流行色算法、八叉树算法和中值切分法, 而  $K$  均值聚类和模糊  $C$  均值聚类是聚类算法的代表。

颜色量化是多项式复杂程度的非确定性(NP)问题, 意味着很难在有限的时间里找到最优解。 $K$  均值算法是一种有代表性的颜色量化算法, 但过分依赖于初始条件, 很容易收敛于局部最优。本文首先利用一种改进的  $K$  均值算法——二分  $K$  均值算法<sup>[4]</sup>进行粗略的颜色量化, 将颜色减少为  $2c$  种, 然后利用基于加权距离的谱聚类进行再次量化。由于二分  $K$  均值聚类不太依赖初始条件、收敛速度快, 且改进的谱聚类算法集成了与像素个数有关的加权距离, 使得新方法在运算速度和量化质量上都取得了不错的结果。像“万绿丛中一点红”这样的图像, 一般的方法很容易出现错划分, 而该方法有效地解决了这一问题。

收稿日期: 2007-07-16; 改回日期: 2007-07-27

第一作者简介: 谷瑞军(1979~), 男, 江南大学信息工程专业博士研究生。主要研究方向为模式识别、数据挖掘、图像处理。E-mail: grj79@

## 2 算法实现

对于调色板的设计,采用由粗到精的策略实现:首先使用二分  $K$  均值聚类将  $N$  种颜色量化为  $2c$  颜色;然后采用改进的谱聚类最终量化为  $c$  种颜色。

### 2.1 使用二分 $K$ 均值算法粗略量化

$K$  均值算法是颜色量化的代表性算法,思路如下:首先定义一个准则函数,并随机选择  $c$  种初始颜色作为簇的质心;然后根据每个像素点与簇质心的距离,将像素点划分到距离最近的簇中;再重新计算每个簇的质心。此过程不断重复,直到准则函数最小或簇质心不再变化为止。最终的簇质心就是调色板的颜色。 $K$  均值算法实现简单,但非常依赖簇的初始质心,很容易陷入局部最优,而且对数据点分布形状也有要求,对呈非椭球形分布的数据往往无能为力。模糊  $C$  均值也有类似的问题。

二分  $K$  均值算法<sup>[4]</sup>是一种改进的  $K$  均值算法,有效地解决了前两个问题。它基于一种简单想法:为了得到  $c$  个簇,将包含所有点的集合分成两个簇;然后根据某种准则选取其中一个,使用  $K$  均值算法分裂成两个簇;如此下去,直到产生  $c$  个簇。二分  $K$  均值算法集成了层次聚类的思想,不易受初始质心的影响,计算简单,而且每一次分裂,都是对整个数据集的某个子集进行二划分,所以比  $K$  均值算法的速度快得多。具体可参见文献[5]。

假定对一幅大小为  $m \times n$  的图像量化为  $c$  种颜色,图像表示为矩阵  $X = [x_1, \dots, x_m]^T$ ,其中,  $x_i$  表示 RGB 颜色空间中的像素点,记作  $x_i = [r_i, g_i, b_i]$ 。用二分  $K$  均值设计包含  $2c$  种颜色的调色板,具体算法如下:

#### 算法 1 颜色的粗略量化

输入:图像  $X = [x_1, \dots, x_m]^T$

输出:调色板  $C = [c_1, \dots, c_{2c}]^T$  和归属关系  $Y = [y_1, \dots, y_m]$

- (1) 初始化簇表为  $C_T = \emptyset$ ;
- (2) 把所有像素点看成一个簇  $clust(1)$ , 加入到簇表  $C_T = \{clust(1)\}$ ;
- (3) 从簇表中选择一个待分裂的簇  $clust(t)$ ;
- (4) for  $i = 1$  to 实验次数
- (5) 使用  $K$  均值划分为 2 个簇,并计算 SSE(离散度和);
- (6) end for
- (7) 选择 SSE 最小的两个簇,并加入到簇表;

(8) 如果  $size(C_T) < 2c$ , 则转到步骤 3;

(9) 根据簇表生成隶属关系  $Y$ , 并计算簇的平均颜色,得到调色板  $C$ 。根据像素点的归属关系和颜色质心进行颜色替换。

其中,  $c_i$  表示簇的质心,即簇的平均色,记作:  $c_i = [r_i, g_i, b_i]$ 。  $Y$  为归属关系向量,  $y_i = m$  表明第  $i$  个像素的颜色应量化为  $c_m$ 。步骤 3 选取待分裂簇的方法有很多种,可以选择包含点最多的簇,或者选择具有最大离散度的簇,也可两者结合。本文以离散度作为选择依据。

### 2.2 使用加权的谱聚类算法再次量化

在使用二分  $K$  均值粗略量化以后,下面采用谱聚类将  $2c$  种颜色减少为  $c$  种颜色。

谱聚类算法<sup>[6]</sup>本质上是利用亲密矩阵的特征向量进行聚类。先给出两个定义:

**定义 1** 亲密矩阵。图  $G = (V, E)$  为加权无向图,节点  $i$  和  $j$  间的距离为  $d(i, j)$ , 则  $A_{ij} = \exp(-d(i, j)/\sigma^2)$  为相似度,其中,  $d(i, j)$  可为各种距离度量。由  $A_{ij}$  组成的对称矩阵  $A$  称为亲密矩阵。

**定义 2** 最小割集。图  $G = (V, E)$ , 图  $G$  被划分为两部分  $P, Q$ , 且有  $P \cup Q = V$ ,  $P \cap Q = \emptyset$ , 将图  $G$  划分为  $P, Q$  两部分的代价为

$$cut(P, Q) = \sum_{u \in P, v \in Q} w(u, v)$$

使代价最小的划分  $(P, Q)$  即为图  $G$  的最小割集。

谱聚类算法的思路如下:首先根据某种度量计算亲密矩阵  $A$ , 根据不同的割集准则构造拉普拉斯矩阵  $L$ ; 然后计算  $L$  的本征值和本征向量,由最大的  $c$  个本征值所对应的本征向量构成矩阵;最后,可假定矩阵的每一行为  $c$  维空间中的数据点,采用普通的聚类算法对数据集进行聚类。

亲密矩阵是谱聚类的关键。对于距离,一般采用欧氏距离。但 RGB 不是均匀的颜色空间,所以一般采用加权距离,即

$$d(i, j) = \sqrt{w_r(r_i - r_j)^2 + w_g(g_i - g_j)^2 + w_b(b_i - b_j)^2} \quad (1)$$

其中,各分量的权值一般取 3、4、2。

考虑到图像千差万别,现有的颜色量化算法对一些特殊图像表现很差。比如像“万绿丛中一点红”这样的图像,虽然红色是这幅图像很重要的颜色,但因包含的像素少很容易被错划分,一般的算法都很难保留住红色。如果相似度  $A_{ij}$  只考虑颜色质心间的距离,而不考虑到颜色的重要程度,则也会出

现错划分。为此,设计了一种可调节权重的加权距离  $d_w(i, j) = w(i, j) * d(i, j)$ , 可以根据需要保留重要颜色(细节), 或侧重层次感(整体)。权函数为

$$w(i, j) = \exp(\beta |h_i - h_j| / \max(h_i, h_j)) \quad (2)$$

其中,  $h_i$  表示颜色  $i$  包含的像素数, 调节因子  $\beta \geq 0$ 。容易得出  $w \in [1, e^\beta]$ 。下面给出权函数  $w$  的两条性质:

**性质 1** 保留层次感。对于层次感比较强的图像, 粗略量化后的颜色相近, 包含的像素数也相近,  $|h_i - h_j| \rightarrow 0$ , 这时无论  $\beta$  如何取值,  $w$  都接近于 1, 保留了原谱聚类算法的优良性质。

**性质 2** 突出细节。如果图像包含重要的颜色(涉及像素少, 但与其他颜色差别大),  $|h_i - h_j| / \max(h_i, h_j) \rightarrow 1, w \rightarrow e^\beta \geq 1$ , 即增大了重要颜色与其他颜色间的距离, 被“孤立”起来, 更容易被保留。

理论上  $\beta$  值越大, 细节越突出, 但太大则也影响图像的层次感, 试验表明可以取 0 到 1 之间。当  $\beta = 0$  时,  $w = 1$ , 即标准的谱聚类算法。基于加权谱聚类的颜色量化算法如下:

**算法 2** 颜色的再次量化

输入: 调色板  $C = [c_1, \dots, c_{2c}]^T$  和归属关系  $Y = [y_1, \dots, y_m]$

输出: 最终的调色板  $C_f = [c_{f_1}, \dots, c_{f_{2c}}]^T$  和归属关系  $Y_f = [y_{f_1}, \dots, y_{f_m}]$

(1) 根据隶属关系, 统计每种颜色包含的像素数, 记作  $h_i, i = 1, \dots, 2c$ ;

(2) 计算任意两种颜色质心间的权重:

$$w(i, j) = \exp(\beta |h_i - h_j| / \max(h_i, h_j))$$

(3) 计算颜色中心间的亲密矩阵  $A: A_{ii} = 0$ , 当  $i \neq j$  时,

$$A_{ij} = \exp(-w(i, j) * |c_i - c_j|^2 / 2\sigma^2);$$

(4) 计算拉普拉斯矩阵  $L, L = D^{-1/2} A D^{-1/2}$ , 其中,  $D$  是对角化矩阵,  $D_{ij} = S_j A_{ij}$ ;

(5) 计算  $L$  的最大的  $c$  个本征值 ( $\lambda_1 = \dots = \lambda_c$ ) 和相应的本征向量 ( $\xi_1, \dots, \xi_c$ ), 按照本征值的大小顺序将相应的本征向量组成矩阵  $S, S = [\xi_1, \dots, \xi_c]$ ;

(6) 对矩阵  $S$  中的每一行进行单位化处理, 即将行向量转变为单位向量, 得到矩阵  $S'$ ;

(7) 将  $S'$  中的每一行视为  $R^c$  空间中的一个点, 使用  $K$  均值算法, 划分为  $c$  个簇, 如果  $S'$  矩阵中的第  $i$  行属于第  $j$  个簇, 则  $c_i$  也属于第  $j$  个簇;

(8) 隶属关系  $Y$  更新为  $Y_f$ , 颜色中心  $C$  更新为  $C_f$ 。

谱聚类算法克服了  $K$  均值算法只适合椭球形数据的缺点, 适合各种形状的数据集, 因此, 划分的簇和人的视觉理解更为吻合。但是亲密矩阵的计算和特征分解计算复杂度很高, 如单纯使用谱聚类对所有的像素点进行聚类, 是不现实的。这也是将调色板的设计划分为两个阶段的主要原因。

### 2.3 算法实现

本文的颜色量化方法是一个由粗到精的两阶段算法 (Two-phase color quantization, TPCQ), 具体步骤如下:

(1) 使用算法 1 将图像量化为  $2c$  颜色;

(2) 使用算法 2 进一步量化为  $c$  种颜色;

(3) 根据归属关系和颜色质心进行颜色替换。

基于二分  $K$  均值聚类的算法 1 使得 TPCQ 稳定且快速, 加权距离使得 TPCQ 可以在突出重要颜色的前提下, 尽可能保留层次感。

## 3 实验和分析

为了验证 TPCQ 的有效性, 选择人工合成图像、标准图像和特殊图像各一幅, 分别使用八叉树、 $K$  均值和 TPCQ 进行实验。

图 1(a) 是大小为  $256 \times 256$  的人工合成图像, 主要有 4 种颜色: 2 种蓝色和 2 种绿色。4 色的量化结果如图 1(b) ~ (d) 所示。八叉树和 TPCQ 方法均能找到 4 种正确的颜色, 且比较稳定。而  $K$  均值方法则很容易得到错误的结果, 如图 1(c) 所示, 原因就是初始的簇质心选择不当。而 TPCQ 的稳定性正是缘于二分  $K$  均值聚类不太依赖初始的簇质心。

图 2(a) 是大小为  $512 \times 512$  的标准 Mandrill 图像, 16 色的量化结果如图 2(b) ~ (d) 所示, 可以看出, 对于普通图像, 八叉树算法的量化结果最差, 而  $K$  均值方法和 TPCQ 的量化结果相差不大。图 3 为同样实验条件下的速度比较, 可以看出, 当颜色数超过 16 色时, TPCQ 的计算速度明显快于  $K$  均值算法。

图 4(a) 是一幅特殊图像, 荷花的颜色很突出, 但包含的像素数很少, 图 4(b) ~ (d) 为 3 种方法的 8 色量化结果。可以看出, 八叉树法和  $K$  均值法均出现了错误划分, 即使运行多次也一样。原因在于红色的像素数太少, 而被归并到相近的颜色中。TPCQ 则给出了满意的结果, 这正是加权函数的作用, 使得红色与其他颜色间的距离加大, 被“孤立”起来, 从而变得更加重要, 被保留下来。

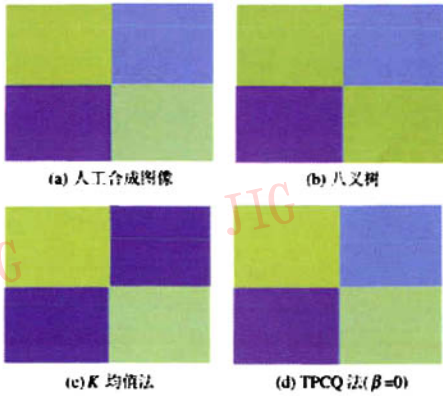


图 1 3 种方法的 4 色量化结果比较

Fig.1 4-color quantization comparison of three methods on artificial image

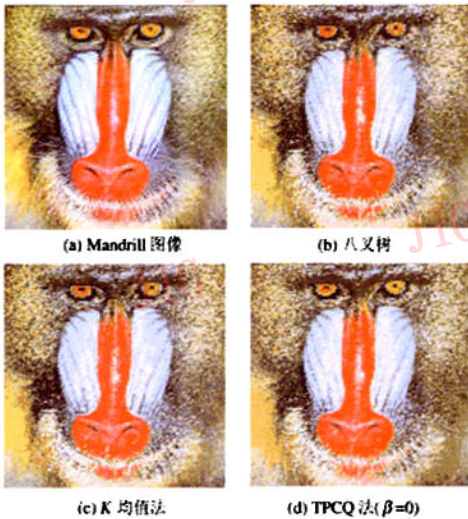


图 2 3 种方法的 16 色量化结果比较

Fig.2 16-color quantization comparison of three methods on Mandrill image

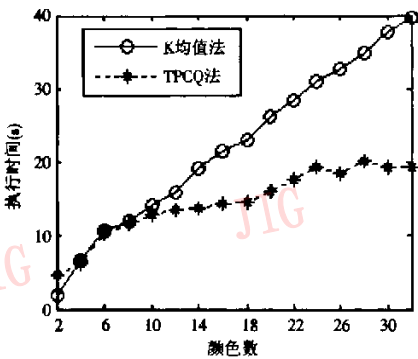


图 3 K 均值和 TPCQ 的运行时间比较

Fig.3 Speed comparison between K-means and TPCQ

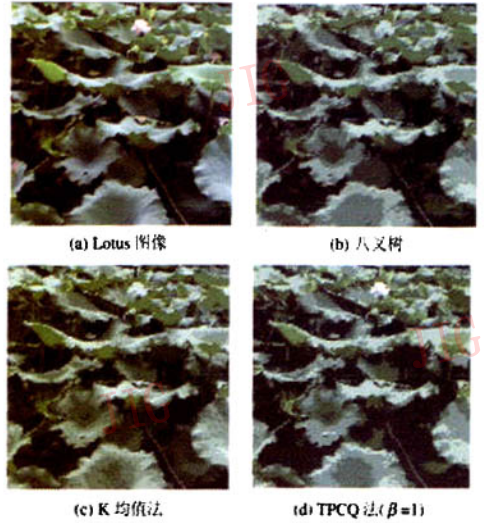


图 4 3 种方法的 8 色量化结果比较

Fig.4 8-color quantization comparison of three methods on special image

### 4 结 论

基于二分 K 均值和谱聚类,提出了一种两阶段的颜色量化算法 TPCQ。首先利用高效、稳定的二分 K 均值聚类进行粗略量化,然后使用基于加权距离的谱聚类进行再次量化。两者的结合使得新算法在运算速度和量化质量上都取得了不错的结果,像“万绿丛中一点红”这样的特殊图像,一般的方法很容易出现错划分,而该算法有效地解决了这一问题。

### 参考文献 (References)

- 1 Fan J, Wu J, Liu Y. A novel color image segmentation algorithm based on vector quantization and region growing[J]. Journal of Image and Graphics, 2005, 10(9): 1079 ~ 1081. [范静辉, 吴建华, 刘晔. 基于矢量量化和区域生长的彩色图像分割新算[J]. 中国图象图形学报, 2005, 10(9): 1079 ~ 1081.]
- 2 Kanjanawanishkul K, Uyyanonvara B. Novel fast color reduction algorithm for time-constrained applications [J]. Journal of Visual Communication & Image Representation, 2005, 16: 311 ~ 332.
- 3 Zhou B, Shen J, Peng Q. An adjustable algorithm for color quantization[J]. Pattern Recognition Letters, 2004, 25: 1787 ~ 1797.
- 4 Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques[A]. In: KDD Workshop on Text Mining[C], Boston, MA, USA, 2000: 109 ~ 111.
- 5 Savaresi S M, Boley D. A comparative analysis on the bisecting K-means and the PDDP clustering algorithms [J]. Intelligent Data Analysis, 2004, 8(4): 345 ~ 362.
- 6 Ng A, Jordan M, Weiss Y. On spectral clustering: analysis and an algorithm[A]. In: Proceedings of NIPS 14[C], Cambridge, MA, USA: MIT Press, 2001: 849 ~ 856.