一种用于彩色图象量化的快速聚类算法

凌 玲
(广东工业大学工程与计算机图学教研室，广州 510090)

摘 要 为了对彩色图象进行有效地压缩处理，提出了一种基于模式识别技术的图象量化新算法(FSCAMMD)，该算法首先把彩色图象中的颜色样本归为一类，并采用最大频度与类内最小距离最大相结合的方法选取初始类代表点——初始值优选法；然后采用欧氏距离聚类准则及重心法，求得新聚类域中心的向量值，从而得到了令人满意的量化效果. 该算法不仅克服了 SCA 算法对聚类中心初始值选取的不足，较大幅度地减少了彩色图象量化后的总方差以及颜色失真度，而且较好地解决了重建彩色图象的整体层次与局部细节之间的矛盾，其量化效果优于 SCA 和其他一些聚类量化算法.

关键词 聚类分析 图象量化 图象压缩 统计
中图法分类号：TP391.41 **文献标识码**：A **文章编号**：1006-8961(2001)08-0771-04

A Fast Clustering Algorithm for Color Images Quantization

LING Ling
(*Dept. of Eng. and Computer Graphics, Guangdong Univ. of Tech., Guangzhou* 510090)

Abstract A new algorithm for color image quantization based on the pattern recognition technology is proposed in this paper. First, the color samples in a color image are grouped together, and the initial representative-points of the categories are chosen based upon a method of combining maximum frequency degree with maximizing minimum discrepancy, that is , an optimum seeking method of initial value of clustering center. Then both the clustering criteria of Euclidean distance in clustering analysis and the gravitational center method in mechanics are used to determine the vector values of the new clustering-region centers, and the satisfying clustering effects can de obtained. This is a fast statistical clustering algorithm based on maximizing minimum discrepancy (FSCAMMD). The presented algorithm can overcome the shortcomings of the seeking method of initial value of the clustering center of SCA algorithm. Both the total mean square deviation and lack fidelity of images quantized by the present algorithm have a relatively big reduction and the effect of color image equalization is better than that of SCA algorithm and other clustering algorithms.

Keywords Clustering analysis, Image quantization, Image compression, Statistics

0 引 言

在多媒体应用和研究领域的诸多待解决问题中，彩色图象的处理是最迫切需要解决的问题之一. 由于互联网的迅速发展，24bit 或更高位真彩色图象的压缩存储和传输问题显得尤为突出，为了有效地进行图象压缩，需要发展通用、可行的处理算法. 量化技术是图象压缩处理的关键技术之一，而调色板选取是量化技术的重要分支[1]，它可以使重建图象在失真度最小的情况下，对彩色图象进行压缩处理.

借鉴模式识别中的聚类分析方法来进行量化，是一种较好的量化方法. 聚类分析是一种数学分析方法，它的数学基础是数理统计的多元分析方法，即它是一种根据样本点很强的内在相似性，把数据表示成一些群、簇的数据描述方法，从而可用一个准则函数，使同一类的向量靠得更近些，以便把它们聚合到一起. 迄今为止，人们已经提出了许多聚类量化算

收稿日期：2001-02-12；改回日期：2001-05-28

法，如经典的 LBG 算法[2]、K 均值算法[3]、ISODATA 算法[4]，FORGY 算法[5]等. 事实上，上述算法均为迭代算法，用其对一般的彩色图象进行量化处理均能得到较好的量化效果，但时空开销大，且对初始聚类中心的选取非常敏感，特别是对色彩分布不均匀的图象，量化效果仍不够理想，因而重建图象整体层次与局部细节之间的矛盾并未得到较好的解决. 另外，由于彩色图象量化不可避免地存在着某些偏差，因此如果要使重建的图象具有丰富的色彩层次，则会失去图象中某些关键细节. 例如一幅绿色草原中一朵红花图象，若要保持原图象整体的绿色层次感，必会使重建图象中红花的红色点着色不正；反之，若要保留红花的着色，则必然会失去原图象整体的绿色层次感. 又如某些图象的平滑部分，由于人眼的视觉特性，对该部分的敏感程度较低，若对该区域进行精细的量化，则反而会造成调色板项的浪费. 因此，如何使得重建图象既能保持丰富的色彩层次，又能使局部的关键细节不被丢失，以及如何根据人眼的视觉特性对图象的各个区域进行适当的量化等问题，迄今未有好的解决方法. 较理想的量化算法应当把误差“平均地”散布在整幅图象中，以至人眼不易观察到量化前后图象的偏差. 要解决以上问题，需要考虑许多因素，如色度空间、图象内容、观察条件、观察者的经验以及美学判断等. 尽管很多研究者讨论了诸多因素，但在这方面还有许多工作要深入.

文献[6]基于聚类原理以及概率论，提出了一种用于彩色图象量化的统计聚类算法(SCA 算法). 其基本思想是：首先在欧氏空间中，采用直方图的峰值搜索法选取初始聚类中心，然后再把所有的颜色样本按最小距离准则向初始聚类中心聚合，以得到若干个集团；最后再利用重心法原理来计算各集团的质心颜色，并将其作为各集团的代表色，以达到图象量化的目的. 由于该算法在选取聚类中心初始值上着重考虑了具体图象色彩的使用频度，从而最大限度地保持了原图象主色调的层次，故比 K 均值算法、ISODATA 算法的量化效果有较大的改善. 但该方法也存在明显的不足，如重建图象的某些局部关键细节较为模糊(颜色偏差较大，正如上面所说的那样，出现红花的红色着色不正)，甚至丢失. 因此，该算法重建的图象质量不够理想，图象的整体层次与局部细节之间的矛盾仍较为突出.

从聚类分析原理可知，由于初始聚类中心的选择是直接影响聚类结果的重要因素之一，而且对同样的模式样本，若选择不同的初始值，可能会得到完全不同的聚类结果，甚至会导致重建图象的严重失真，因此，合理选择初始聚类中心是解决最佳划分问题的关键. 本文以 SCA 聚类算法为基础，对该算法聚类中心初始值选取的方法作了有效的修正，提出了一种快速聚类算法(FSCAMMD)，其量化效果均比以上算法有较大幅度的改善.

1 FSCAMMD 聚类算法

为了使算法少一些主观因素，应尽可能多地保留图象的原始信息，为此在参考了图象色彩频度的前提下，采用了类内最小距离最大的方法来选取初始聚类中心. 其基本步骤是：在 m 维欧氏空间中(对彩色图象，$m=3$)，对拥有 L 种颜色的彩色图象进行 N 级量化时，先把 L 个颜色样本(记为样本集合 $z=\{z_i^0 \mid i=1,2,\cdots,L\}$)归为一类，并指定其中 $P(1\leq P\leq N)$ 个样本点为初始类的代表点，记为集合 $z^1=\{z_k^1 \mid k=1,2,\cdots,P\}$，被指定的样本点应是 z^0 样本集合中频度最高的前 P 个样本点；然后在其余的样本集合 $z^2=\{z_k^2 \mid k=1,2,\cdots,L-P\}$ 中，再找出与初始类代表点最小距离最大的样本点作为补充的类代表点；依此类推，直到找出 N 个类代表点为止，最后把它们作为初始聚类中心. 其算法过程如下：

(1)统计整幅图象的颜色样本，得到颜色直方图 $h(i)$

与颜色样本 z_i^0 对应的频度为 $h(i)(i=1,2,\cdots,L)$.

(2)按颜色样本频度选取 P 个初始类代表点.

根据直方图选取 $h(i)$ 值最大的前 P 个样本点作为初始类代表点，构成集合 z^1. 对多幅图象进行的实验结果表明：对于一般的彩色图象，当取 $P=N/2$ 时，即 $z^1=\{z_k^1 \mid k=1,2,\cdots,N/2\}$，图象量化后的总方差及颜色失真度误差最小.

(3)采用类内最小距离最大的方法选取 $N-P$ 个初始类代表点.

Step 1 令 $\|z^1\|=k$，$\|z^2\|=f$

Step 2 计算 z^2 中所有的样本点与 z^0 中各样本点 z_i^2 的最小距离，并进行比较，然后取其中最大者，记为 $D_j(j\in\{1,2,\cdots,f\})$，它所对应的样本点记为 z_j^2，其中

$$D_j=\bigvee_{n=1}^{f}\left(\bigwedge_{m=1}^{k}D_{mn}\right) \tag{1}$$

式中，∨和∧(模糊数学中的运算符号)分别表示取大运算和取小运算，而

$$D_{mn}=\|z_n^2-z_m^0\|$$
$$=\sqrt{(r_n-r_m)^2+(g_n-g_m)^2+(b_n-b_m)^2} \quad (2)$$
$$n=1,2,\cdots,f;m=1,2,\cdots,k$$

式中，$z_n^2\in z^2$，$z_m^0\in z^0$，r_n,g_n,b_n 分别是 z_n^2 样本点的三色分量，r_m,g_m,b_m 分别是 z_m^0 初始类代表点的三色分量.

Step 3　$z^1=z^1\cup\{z_j^2\},z^2=z^2-\{z_j^2\}$　(3)

Step 4　若$\|z^1\|<N$，则转 Step 1.

(4)按欧氏距离聚类准则，把图象中所有的颜色样本，向 N 个初始聚类中心聚合，从而得到 N 个聚类域，记为 $s_1,s_2,\cdots,s_N$.

(5)计算各聚类域的代表色(质心颜色).

借鉴力学中的重心(或质心)计算法则，则每一聚类域代表色 z_k^3，可由下式求得

$$z_k^3=\frac{1}{m_k}\sum_{z_i^0\in s_k}z_i^0\times h(i)$$
$$i=1,2,\cdots,L;k=1,2,\cdots,N \quad (4)$$

式中，$z_k^3\in z^3$，$z^3=\{z_k^3|k=1,2,\cdots,N\}$ 为各聚类域代表色的集合，m_k 为 s_k 聚类域中包含的所有颜色样本总数. 至此，便可以从拥有 L 种颜色的彩色图象中得到 N 种最佳的代表色，从而得到 N 级的量化效果.

2　理论与实验分析

在聚类中心初始值的选取中，采用了最大频度与类内最小距离最大相结合的方法，并从使聚类失真度最小的角度出发. 显然这种预置在理论上是合理的. 这样不仅可以使它们一开始就接近最终的聚类中心，而且可以得到一个类内团紧密、类内最大距离最小的最佳划分. 也正因为如此，本算法不仅比 SCA 算法对图象量化后的总平均误差获得了较大的优化，同时也有效地减小了图象中局部颜色的偏差.

为了检验本算法的实际处理效果，分别用 SCA 算法和本文算法对一幅颜色分布较均匀的风景图象(图版 I 图 1(a))和一幅颜色分布极不均匀的的花卉图象(图版 I 图 2(a))(均为未经压缩处理的 256 色彩色图象)进行了 16 级量化处理实验，其得到量化后的图象分别如图版 I 图 1(b)、(c)以及图版 I 图 2(b)、(c)所示. 本文还对各种级别下的量化误差做了进一步的比较，其结果如表 1 和表 2 所示，其中，颜色失真度的定义见文献[7].

表 1　FSCAMMD 算法与 SCA 算法对图版 I 图 1(a)的量化误差比较

量化级数	颜色失真度误差		总平均误差	
	SCA 算法	FSCAMMD 算法	SCA 算法	FSCAMMD 算法
128	1.007 30	0.994 39	2.508 24	2.020 66
64	2.328 81	1.694 06	5.502 86	3.900 16
32	3.258 00	2.475 26	8.165 11	6.645 09
16	5.842 49	4.034 55	22.645 14	10.386 01

表 2　FSCAMMD 算法与 SCA 算法对图版 I 图 1(a)的量化误差比较

量化级数	颜色失真度误差		总平均误差	
	SCA 算法	FSCAMMD 算法	SCA 算法	FSCAMMD 算法
128	0.552 37	0.679 92	1.236 43	1.276 51
64	1.545 10	1.420 72	3.195 38	2.592 95
32	2.906 53	2.035 96	5.160 14	4.120 45
16	3.685 68	2.705 32	7.158 05	6.540 67

对比图版 I 图 1(a)、(b)、(c)可见，由于该图象在色度空间中分布较均匀，故本算法量化效果比 SCA 算法(图版 I 图 1(b))有明显的改善，且令人满意. 尤其是对图版 I 图 1(a)图象进行 16 级量化时，由于本算法的总平均误差比 SCA 聚类算法减小了一倍多(表 1)，因此重建图象的颜色数虽然只有 16 种，但肉眼几乎感觉不到明显的差别，且噪声点的处理也趋于理想.

对比图版 I 图 2(a)、(b)、(c)也不难发现，即使图象颜色分布极不均匀，本算法也可以得到较理想的量化效果. 虽然 SCA 算法仍能较好地保留了玫瑰花瓣的红色层次(图版 I 图 2(b))，但局部区域的关键颜色，如玫瑰花托的颜色却损失较大，失去了原有的绿色，导致出现了人眼觉察到的明显偏差. 而本算法则不仅保留了原图象整体的颜色层次，而且避免了玫瑰花托的绿色不致于被严重丢失，从而较好地解决了 SCA 算法以及其他一些聚类算法中重建图象的整体层次与局部细节之间的矛盾. 误差统计结果分析表明(表 2)，本算法的量化总平均误差和颜色失真度误差大多数均小于 SCA 算法. 从表 2 中还可以发现，在 128 级量化时，SCA 聚类算法颜色失真度和总平均误差小于本算法. 这表明，当原图象颜色分布极不均匀时，初始类代表点的选择以及数量对量化效果的影响较大，这是本算法有待解决的问

题.此外,本算法对噪声点的控制还可以作进一步深入地研究.

3 结束语

本文在 SCA 聚类算法的基础上,提出了一种 FSCAMMD 快速聚类算法.该算法克服了 SCA 算法中初始聚类中心选取的不足,即采用频度最大与类内最小距离最大的方法来选取初始聚类中心,仅通过无迭代的一次聚合运算,即可得到令人满意的量化效果.理论分析与实验结果均表明,本文算法不仅保持了 SCA 聚类算法的简洁、速度快等特点,而且对全局的优化效果以及局部颜色的控制均比 SCA 聚类算法以及其他一些聚类量化算法有显著地提高.

参 考 文 献

1 Sharma Ganrav, Joel Trussell H. Digital color imaging[J]. IEEE Transaction on Imaging Processing, 1997, 6(7): 901～932.

2 Linde Y, Buzo A, Gray R M. An algorithm for vector quantizer design[J]. IEEE Trans. on Commun, 1980, COMM-28: 84～89.

3 Kamel M S, Selim Z. New algorithms for solving the fuzzy clustering problem[J]. Pattern Recognition, 1994, 27(3): 421～428.

4 Venkateswarlu N B, Raju PSVSK. Fast ISODATA clustering algorithms[J]. Pattern Recognition, 1992, 25(3): 335～342.

5 戚飞虎,周源华等译. 模式识别与图象处理[M]. 译自 Hamolbook of Pattern Recogocition and Image Processing, ACADEMIC Press, Inc, 1986, 上海: 上海交通大学出版社, 1990.

6 凌玲. 彩色图象量化方法的研究[J]. 华南理工大学学报, 2000, 28(1): 81～84.

7 凌玲, 凌卫新. 彩色图象量化的一种新聚类算法[J]. 广东机械学院学报, 1996, 14(2): 58～621.

凌 玲 1960 年生,1995 年获华南理工大学机械电子工程系硕士学位,现任广东工业大学工程与计算机图形学教研室讲师.主要研究方向为小波分析、图象处理、模式识别、计算机图形学等.

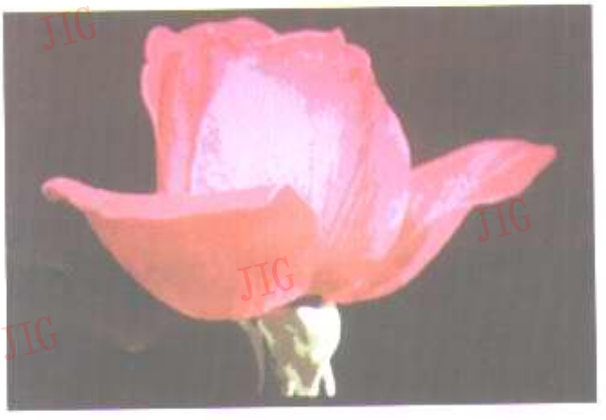
(a) 原图

(b) 原图

(c) SCA 算法重建的图象(N=16)

(c) SCA 算法重建的图象(N=16)

(d) FSCAMMD 算法重建的图象(N=16)



(d) FSCAMMD 算法重建的图象(N=16)

图 1 用不同算法重建的风景图象

图 2 用不同算法重建的花卉图象