

# 基于颜色直方图的图象检索

伯晓晨 刘建平

(国防科技大学自动控制系, 湖南长沙 410073)

**摘要** 针对基于颜色直方图的静态图象检索, 提出了自动构造全局参考颜色表和自带参考颜色表的算法, 解决了 Swain 算法中存储量与计算量过大的问题。同时给出了一个通过分层聚类构造二叉树式分层索引数据结构的方法。

**关键词** 基于内容检索 直方图相交 参考颜色表 聚类二叉树

## 0 引言

信息时代的到来使得人们越来越多地接触到大量的图象信息, 有效地组织、管理和检索大规模的图象数据库成为迫切需要解决的问题。传统的数据库检索采用基于关键词或描述性文本的检索方式, 查询时需要指明文本特征, 要求用户对文本特征的描述具有一定的准确性和规范性。由于图象内容的丰富内涵以及人们对图象内容进行抽象时的主观性, 使得这种方法往往不能满足人们的需要。

基于内容的检索(Content Based Image Retrieval, CBIR)不同于传统的检索手段, 它融和了图象理解技术, 从而可以提供更加有效的检索途径。目前国外许多机构都在进行相关的研究, 并推出了以IBM的QBIC、Berkley的Chabot和MIT的PhotoBook为代表的一系列成功的产品。

基于内容检索的3个突出特点是:

- (1) 用于检索的是反映图象内容并与图象存储在一起的各种量化特征。
- (2) 使用基于相似性度量的近似查询。
- (3) 大多采用示例查询(Query By Example)方法。

针对静止图象的检索, 可使用的特征有颜色、纹理、草图和形状等等。颜色在传统的图象理解中未能

受到普遍重视, 然而相对于几何特征而言, 颜色具有与生俱来的旋转不变性(rotation-invariance)和尺度不变性(scale-invariance)。在基于内容的图象检索中, 颜色是使用最广泛的特征之一。图1显示了一幅商标图象其红色(Red)通道中直方图的旋转不变性。

基于颜色检索的基本思想是将图象间的距离归结为其颜色直方图间的差距, 从而图象检索也就转化为颜色直方图的匹配。1990年 Swain 和 Ballard<sup>[1]</sup>提出了基于直方图相交(histogram intersection)的色彩索引(color indexing)算法, 成为这一方面的奠基性工作。不过 Swain 算法要存储和计算整个颜色空间内的直方图, 计算量与存储量过大。为此, Mehtre 等提出了用参考颜色表(reference color table)方法来压缩直方图<sup>[2]</sup>。然而 Mehtre 的算法依赖于存储图象色彩分布的先验知识, 随着图象库中数据的增删, 参考颜色表会变得越来越不精确。

另外对于大量的图象数据来说, 如果要用线性表的结构来组织, 检索时间将是线性增长的, 而若采用树状分层结构就必须解决从原始特征矢量到结点位置的映射。

在本文中我们采用聚类方法来解决参考色的选择和树状存储结构的构造问题, 并介绍了针对一组商标图象的实验结果。

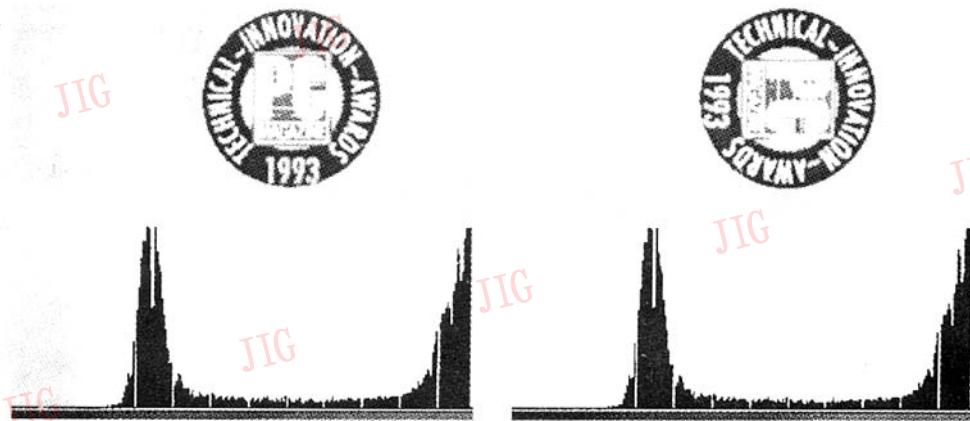


图1 色彩直方图的旋转不变性

## 1 R、G、B 颜色直方图及其差异度量

设示例图象 (Queue Image) 为  $Q$ , 存储图象为  $I$ 。

文献[2]曾经对直方图差异的度量做过最简单的处理,即用亮度均值作为亮度分布的代表。定义 R、G、B 空间中的特征矢量,  $f = (\mu_R, \mu_G, \mu_B)$ ,  $\mu_R, \mu_G, \mu_B$  分别是彩色图象在 R、G、B 3个通道中的亮度均值,则  $Q, I$  之间的距离定义为:

Manhattan 距离:

$$D_{Q,I}^M = |f_Q - f_I| = \sum_{R,G,B} |\mu_Q - \mu_I|$$

Euclidean 距离:

$$D_{Q,I}^E = \sqrt{(f_Q - f_I)^2} = \sqrt{\sum_{R,G,B} (\mu_Q - \mu_I)^2}$$

Swain 和 Ballard 提出的直方图相交是目前使用最广泛的算法。设  $Q, I$  在某颜色通道  $c$  上的颜色直方图分别为  $Q_c(i)$  和  $I_c(i)$  其中  $i \in [1, n]$ ,  $n$  为亮度的水平数,则在通道  $c$  上的直方图距离  $D_{Q,I}$  定义为:

$$D_{Q,I} = \frac{\sum_{i=1}^n \min\{Q_c(i), I_c(i)\}}{\sum_{i=1}^n I_c(i)}$$

为了做到尺度不变,使用归一化的直方图  $I_c(i) = \frac{I_c(i)}{\sum_{i=1}^n I_c(i)}$  及  $\tilde{Q}_c(i) = \frac{Q_c(i)}{\sum_{i=1}^n Q_c(i)}$ , 当  $I, Q$  点数

相同时,有  $D_{Q,I} = \sum_{i=1}^n \min\{\tilde{Q}_c(i), \tilde{I}_c(i)\}$ 。

综合 R、G、B 3个通道,则  $Q, I$  之间的距离为:

$$D_{Q,I} = \sum_{c=R,G,B} \left[ \sum_{i=1}^n \min\{\tilde{Q}_c(i), \tilde{I}_c(i)\} \right]$$

Swain 证明了直方图相交运算的结果反映的是 2 个直方图的街区距离,说明了上述方法的物理含义。另外为了考虑不同颜色之间的相似性,有许多研究者还采用了其它的距离度量方法,如定义

$$D_{Q,I} = 1 - \sqrt{\frac{1}{6} \sum_{c=R,G,B} \left[ \sum_{i=1}^n (\tilde{Q}_c(i) - \tilde{I}_c(i))^2 \right]}$$

## 2 全局参考颜色的选择

使用 R、G、B 3个直方图来表征图象的色彩特征无疑是准确的,但由于要存储和计算每幅图象的原始直方图,所付出的代价也是相当大的。

为了压缩色彩直方图,Mehtre 等人在文献[2]中使用了参考颜色表方法。

在长度为  $n$  的参考颜色表上计算每个图象的色彩直方图,图象中的每一色点被归入相近的参考颜色。图象的特征矢量就定义为这个参考表上的归一化直方图,即  $f = (h_1, h_2, \dots, h_n)$ 。图象间的距离定义

为  $D_{Q,I} = \sum_{i=1}^n \sqrt{(h_i^Q - h_i^I)^2}$ 。可以看出,参考色的选

择是至关重要的,如果没有存储图象的先验知识,参考颜色表也就无从确定。

在对注册商标图象库的研究中,我们发现由于商标设计者的原因,实际中每幅图象所包含的主要颜色是不多的,主要颜色之外的其他颜色大多是由于扫描输入时的色散效应造成的。此外,在 R、G、B 真彩色空间中还存在着大片空白的区域。图2是一个

典型商标图象的 R 通道直方图。

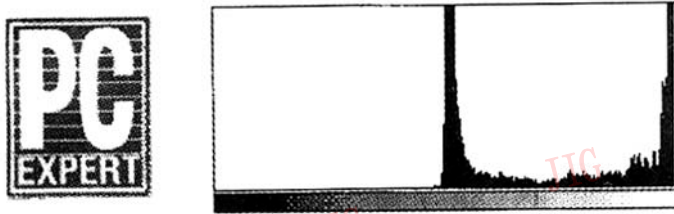


图2 一幅典型商标图象的红色通道直方图

为了在没有先验知识的情况下自动地从存储图象中抽取参考色,我们打破了图象之间的界限,在统计整个库存图象颜色的直方图后,使用聚类方法生成参考色表。

算法一 全局参考色直方图的自动获取

Step1 打破图象间的界限,统计反映整个库存图象色彩分布的联合直方图(三维)

$$f_1 = (h_1, h_2, \dots)$$

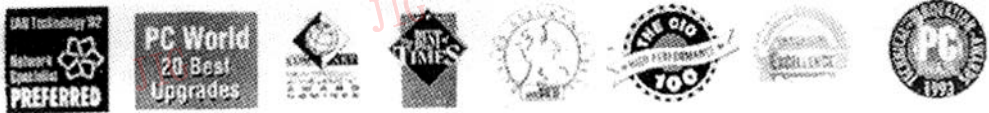
Step2 为消除量化噪声的影响,对直方图进行均值平滑,得到  $f_2 = (h'_1, h'_2, \dots)$ 。

Step3 搜索  $f_2$  得到  $m$  个峰值,按其对应的象素数从大到小排序得侯选序列  $h^{(1)}, h^{(2)}, \dots, h^{(m)}$ 。

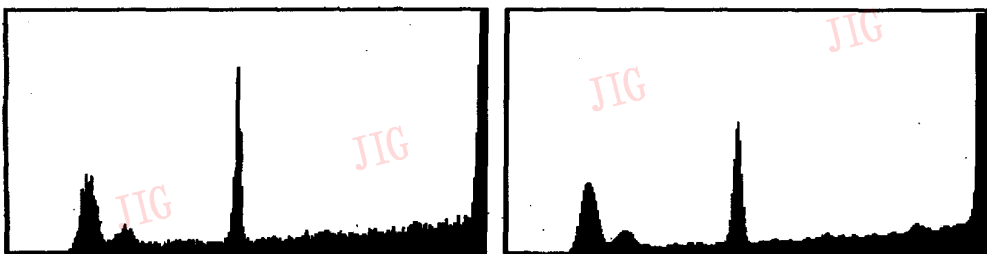
Step4 对侯选序列截尾,去掉点数小于阈值 TH 的颜色,得到长度为  $n$  的参考颜色表  $(C_1, C_2, \dots, C_n)$ 。

Step5 对存储的每一幅图象,以各参考色为聚类中心,用 R、G、B 空间中的  $L_1$  距离和最小距离方法将相近的颜色聚成一类,同时统计参考色表上的分布,得到参考颜色直方图,以此作为图象的特征。这实际上是一个聚色的过程。

图3是8幅商标图象在 R 通道内的联合直方图及平滑处理后的直方图,其中包含了有限个高峰值,色彩分布不均匀的现象是很明显的。



原始商标图象



联合分布直方图

平滑后的直方图

图3 商标图象极其联合分布直方图

实验中我们针对六十余个计算机行业的商标图象进行统计,确定参考颜色表长度为7。

### 3 自带参考颜色表

全局参考颜色表方法的突出缺点是当对数据库有增加或减少图象的操作时要重新计算全局参考颜

色表,计算量随数据库规模的扩大迅速增长。此外,对于那些色彩与大多数图象不同的图象来说,在全局参考颜色表上统计的直方图与原始直方图差距太远,使得检索的准确性难以保证。为此我们提出了一个改进算法,即每个图象根据其自身色彩分布用聚类方法生成自带的参考颜色表,在这个参考颜色表上聚色并重新统计直方图,以此作为图象的特征。这

里唯一固定的只是参考颜色表的长度  $L$ 。

**算法二** (自带参考色直方图的获取):

- Step1 对每幅图象,统计其三维色彩分布直方图  $f_1 = (h_1, h_2, \dots)$ 。
- Step2 对直方图进行均值平滑,得到  $f_2 = (h'_1, h'_2, \dots)$ 。
- Step3 搜索  $f_2$  得到  $m$  个峰值,按其对应的象素数从大到小排序得候选序列  $h^{(1)}, h^{(2)}, \dots, h^{(m)}$ 。
- Step4 对候选序列截尾,只取前  $L$  个颜色,得到长度为  $L$  的参考颜色表  $(C_1, C_2, \dots, C_n)$ 。
- Step5 聚类:以各参考色为聚类中心,用 R、G、B 空间中的  $L1$  距离和最小距离方法将相近的颜色聚成一类,同时统计参考色表上的分布,得到参考颜色直方图。
- Step6 对参考颜色直方图按点数对色彩排序,即依次找出最主要颜色、次主要颜色、次要颜色、...、如果点数相同则按其与其空间原点的距离确定先后。图象的特征定义为顺序排列的颜色属性  $(H_1, H_2, \dots, H_L)$ , 其中  $H_i = \{R_i, G_i, B_i, h_i\}$ , 其中既包括了参考色的 R、G、B 值,又包括了相应的点数  $h_i$  (已归一化)。

此时,图象间的距离定义为:

$$D_{Q,I} = \sum_{i=1}^L \omega_i \times \left[ \sqrt{(R_i^Q - R_i^I)^2 + (G_i^Q - G_i^I)^2 + (B_i^Q - B_i^I)^2} \right]$$

$$\text{其中 } \omega_i = \frac{1}{2} (h_i^Q + h_i^I)$$

可以看出,此时对距离起决定作用的是2幅图象在各个级别上主要颜色的种类差异,而具体的点数只作为加权。使用商标图象所做的实验结果表明,每幅图象只保留3个主要颜色(即  $L=3$ )就可以取得较好的结果。

## 4 二叉树存储结构的实现

衡量一个图象数据库检索能力的标准除了速度外,还有它的可靠性,也就是能否把所有相似的图象都检索出来。有可能检索出来的图象并不重要,但重要的绝对不能遗漏。另外,相似性检索应是在多精度水平上进行的,随着精度的降低应该得到越来越多的匹配结果。

为统一解决速度、可靠性和多精度问题,我们改

进了层次聚类算法,将存储的图象逐步聚类,类别由多到少,最终形成二叉树存储结构。

**算法三** 聚类方法生成二叉树存储结构

- Step1 库存的  $N$  个图象自成一类,共计  $N$  类,记为  $C_1^0, C_2^0, \dots, C_N^0$ , 基于参考颜色直方图计算各类间的距离,得到  $N \times N$  维距离矩阵  $D^0$ 。
- StepK+1 已求得  $D^K$ , 求出  $D^K$  中最小元素,将对应的2类  $C_i^K, C_j^K$  合并成一类,由此得到新的分类  $C_{i+j}^{K+1}, C_{i+j}^{K+1}, \dots$

如此不断重复,每步都合并2类,直至将全部图象聚成一类,形成一个根节点。这样就构成了一个二叉树的结构,树的节点数为  $2N-1$ , 深度小于  $\log_2 N$ 。需要注意的是,将  $C_i^K$  与  $C_j^K$  合并成一类,暂记为  $C_{i+j}^K$ , 要计算新的类别到其它类别的距离,一般可采用的方法有最短距离法、最长距离法、中间距离法、重心法等,详见文献[7]。图4给出了5幅图象聚类成二叉树的过程。

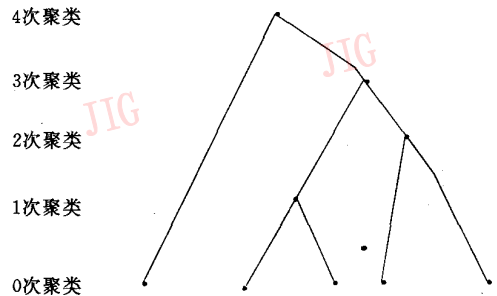


图4 自底向上的分层聚类

对于大规模的图象数据库来说,如果每增加一幅图象都这样聚类,其时间的耗费是难以忍受的,而且也没有充分利用原数据库结构的信息。为此我们在实现图象增、删的过程中只在树状结构的局部通过算法三重新调整图象的位置,而保持绝大部分的数据库结构不变。

当查询者给出一个要检索的例图  $Q$  后,抽取其参考颜色并统计参考色直方图后得到特征矢量,检索从根节点开始,不断向下搜索与其相近的节点,直到找到一个叶节点。而且根据不同的精度要求,还可以将某层某类下的一批图象检索出来。这样不仅满足了速度要求,而且做到了多精度水平的匹配。实践证明这种数据结构是非常有效的。

## 5 结论及后续的工作

在我们使用商标图象所做的实验中,本文中的算法取得了较好的结果,然而并不是说这些算法对所有的图象数据库都适用,它还仅仅是基于内容检索这种新技术的一个雏形。尚未解决的问题包括亮度不变性、颜色特征与其它特征的综和使用等等。在以后的工作中,我们将主要研究如何将颜色特征与形状特征结合起来。

### 参考文献

1 Swain M J, Ballard D H. Color Indexing. *int. J. Computer Vi-*

sion, 1991, 7(1): 11~32.

2 Mehtre B M, Kankanhalli M S. Color Matching for Image Retrieval. *Pattern Recognition Lett.*, 1995, 16: 325~331.

3 Kankanhalli M S. Cluster-Based Color Matching for Image Retrieval. *Pattern Recognition*, 1996, 29(4).

4 白雪生,徐光佑, Jin J S. 基于内容的图象检索及其相关技术的研究. *机器人*, 1997(5).

5 冯玉才,宋恩民. 彩色地图图象的聚色算法研究. *软件学报*, 1996, 7(8).

6 曹莉华,李国辉,胡晓峰. 基于相关度和注意力聚焦的图象内容索引. *国防科技大学学报*, 1997, (2).

7 边肇祺. 模式识别. 北京:清华大学出版社, 1990.



伯晓晨 硕士研究生, 1973年生, 1996年毕业于国防科技大学自动控制系, 研究方向为计算机视觉、图象数据库。



刘建平 教授, 1957年生, “863”智能机器人专家组成员, 研究方向为模式识别与计算机视觉。

## Color Histogram Based on Image Retrieval

Bo Xiaochen, Liu Jianping

(Department of Automatic Control, National University of Defense Technology, Changsha 410073)

**Abstract** Content-Based image retrieval (CBIR), which is the mixture of image understanding and multimedia, is becoming a key technique in next generation large scale intelligent multimedia database. In this paper, we present three algorithms to support color histogram based on retrieval in logo image database. The first two algorithms are used to reduce the complexity in Swain's algorithm, and the third one is used to make binary tree structure for quick retrieval.

**Keywords** Content-based retrieval, Histogram intersection, Reference color table, Cluster, Binary tree

## CA 全球专业服务在 Internet Open 开放论坛上赢得“最佳等级”大奖

Computer Associates (CA)公司的全球专业服务(GPS)机构由于建立了一个 Lotus Domino 应用程序,大大加速了 Key 公司的一个子公司的投资回报,而在 1998 Internet Open 开放论坛上赢得了群件及企业支持方面的“最佳等级”大奖。

GPS 是 Lotus 的特级业务伙伴,在经过了有 60 多个竞争者在 8 个类别方面的激烈竞争之后, GPS 赢得了这个大奖。它的解决方案几乎在所有评估类别(包括用户体验、技术革新和投资回报)方面都获得了最高分。评估标准包括技术/业务革新,用户友好性,用户效率,安全/可靠属性,平台灵活性,视觉效果,以及经济有效性/终端用户业务收益效果。

GPS 的执行副总裁 Chris Wagner 说:“这个奖项证明了 CA 公司的全球专业服务机构的两个最重要的属性——无可匹敌的技术专长和一种清晰的能力。由于今年 Internet Open 开放论坛中的竞争异常激烈,所以我们为能够获得这项荣誉而感到非常高兴”。

GPS 建立的应用程序使 Key 资产管理公司的信息建立和传输更加合理。Key 资产管理公司是 Key 公司的投资顾问部门。它使得批准过程流线化,对公司的知识信息提供了一个中央存储,并将最终文档、文件和出版物传递给企业以及相关的业务伙伴。缩短生产周期,提供一个有效的、低成本的传输机制,以及关闭反馈循环而提高生产率,通过这些,它加速了企业的投资回报。