基于改进颜色聚合向量与贡献度聚类的图像检索算法

张永库[1] 李云峰[2] 孙劲光[1]
(辽宁工程技术大学电子与信息工程学院 葫芦岛 125105)[1]
(辽宁工程技术大学研究生学院 葫芦岛 125105)[2]

摘 要 为了提高图像检索的速度和准确率,通过分析各种聚类算法在图像检索中的缺点,提出了一种新的划分聚类的图像检索方法。首先,在对HSV模型非均匀量化的基础上,利用改进的颜色聚合向量方法提取图像的颜色特征;然后找到符合条件的特征向量作为初始聚类中心,利用分散度与贡献度进行聚类并建立特征索引库;最后根据查询图像的相似度进行检索和排序。实验结果表明,所提算法的查准率和查全率比其它算法均有较大提高。

关键词 HSV模型,颜色聚合向量,分散度,贡献度
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.066

Image Retrieval Algorithm Based on Improved Color Coherence Vectors and Contribution to Clustering

ZHANG Yong-ku[1] LI Yun-feng[2] SUN Jing-guang[1]
(College of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China)[1]
(Institute of Graduate, Liaoning Technical University, Huludao 125105, China)[2]

Abstract In order to improve the speed and accuracy of image retrieval, the drawbacks of image retrieval based on a variety of clustering algorithms were analyzed, and a new partition clustering method for image retrieval was presented in this paper. First, based on the asymmetrical quantization of the color in HSV model, color coherence vectors are introduced as the color feature. Secondly, qualified feature vectors are found as the initial cluster centers, and it clusteres based on the dispersion and the contribution, establishes image feature index library. Finally, it obtains the retrieval and reordered results by the similarity with the retrieval image. By comparing with other algorithms, it is demonstrated that the percentage of precision and recall of proposed algorithm are improved greatly.

Keywords HSV model, Color coherence vectors, Dispersion, Contribution

1 引言

由于多媒体技术的普及,网络技术与数字图像技术的发展,图像资源变得越来越丰富。为了从海量图像中迅速地找到感兴趣的图像,基于内容的图像检索(CBIR)技术得到了广泛应用并且成为研究热点。目前相关研究已发展20多年,产生了许多用于实践的CBIR系统[1],如IBM公司开发的最早商业化QBIC系统、哥伦比亚大学研发的WebSeek系统、麻省理工学院研发的PhotoBook系统、Google的Google Similar Images、百度的百度识图等。

CBIR核心技术包括视觉特征的提取、相似性计算、系统性能评价等。CBIR系统主要利用图像的底层特征对图像进行检索。图像的底层特征主要包括颜色、纹理[2]、形状和空间关系等,颜色是彩色图像最底层、最显著、最重要的特征,是绝大多数基于内容的图像检索经常使用的特征之一,很多基于颜色的图像检索方法已被提出。

聚类是多媒体数据挖掘的重要任务之一,是无监督分类的一种形式,聚类算法在图像检索领域中有着广泛的应用[3-7]。利用聚类算法将特征相似的图像聚合为一类,能够提高检索的速度和精确度,因为其检索过程在类心和某一类内进行,图像检索和匹配的范围较小。在所有方法中,K-means聚类算法应用最广泛,其优点是概念简单、通用、容易实现;该算法检索速度比较高,保证以二次收敛的速度终止。K-means算法主要的缺点是经常终止于一个局部最小值,该算法对初始类心依赖性强,导致检索结果不稳定。近几年,许多人也提出了许多基于改进K-means算法的图像检索算法[8-14]。2010年,V. S. V. S. Murthy等人提出了基于分层和K-means算法的图像检索算法[8],即提取颜色直方图,利用分层聚类算法选取初始类心,然后利用K-means算法迭代优化,虽然图像检索的查准率和查全率有所提高,但是不明显,而且算法的时间复杂度较大。2012年,HO Jan-ming等人提出了一种综合多模块的图像检索算法[11],首次提出将图像分割为若干网格块,然后提取每个网格的颜色值,利用上下文差异直方图(Contrast Context Histogram)提取颜色特征并建立近邻表(Neighborhood Table),利用

到稿日期:2014-04-10 返修日期:2014-06-16 本文受国家科技支撑计划(2013bah12f01)资助。
张永库(1972-),男,硕士,副教授,主要研究方向为图形图像和多媒体、数据处理和数据挖掘,E-mail:godlovelyf@163.com;李云峰(1989-),男,硕士生,主要研究方向为图像检索、计算机图形图像处理;孙劲光(1962-),女,博士,教授,主要研究方向为图形图像处理与人脸识别、数据挖掘。

K-means 方法进行聚类。此方法总体上提高了图像检索的查准率和查全率，但是时间复杂度比较大，在大图像库中进行检索显然不是一个好的方法。2013 年，吕明磊等人提出了一种改进的 K-means 算法的图像检索算法[12]，针对 K-means 算法随机选取初始值导致聚类结果不稳定的特点，提取图像颜色直方图，利用累加和公式选取初始聚类中心。该算法保证了检索结果的稳定性，提高了检索的准确性，但是总体来说图像检索的效果有待提高。2014 年，LIN Chuen-horng 等人提出了一种加速的 K-means 算法[14]，首次考虑了图像库的动态性(图像库中图像数量增多或减少)，利用颜色直方图的离散度函数和 K-means 聚类算法选取图像库中的类心，降低了图像库动态变化时图像检索的时间复杂度，但是该算法的检索精度较差。模糊 C 均值算法[15]的初始聚类中心是利用伪随机数产生的，聚类效果不稳定，在图像种类比较多时，检索效果通常不好。K-modes 聚类算法[16]具有脏数据和异常数据不敏感性，虽然是 K-means 算法的扩展，但计算量要远大于 K-means 算法，不适合大量图像的检索。另外，文献[17,18]也介绍了其它基于聚类的无监督学习的 CBIR 技术。

以上算法还有一个不足之处，即仅仅考虑了类内的相似度而没有考虑类间的分散度，导致其在整体上没有更好的检索效果。本文提出的算法利用改进的颜色聚合向量提取颜色特征，着重考虑了颜色空间分布信息，并且基于贡献度的聚类的迭代特性保证了较低的时间复杂性，而且综合考虑了图像类内的相似度和类间的分散度，提高了图像检索的查准率和查全率。

2 图像检索流程

本文图像检索系统的简要流程如图 1 所示，主要包括 3 方面：首先，收集和分析图像资源，将输入图像放入图像库中，同时提取其特征存入特征库；然后，利用相关算法对特征库中的特征向量进行聚类，将聚类的结果存入特征库；最后，对给定的查询图像进行分析并且提取其特征向量，计算其与聚类中心的相似度，查询图像归属于与其相似度最大的类，然后计算查询图像与该类中所有特征向量的相似度大小，按照降序的方式输出该类所属图像。

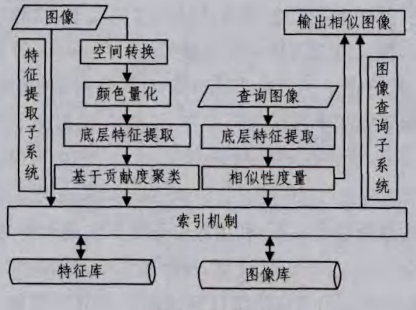

图 1 系统流程

3 特征提取

3.1 HSV 颜色的非均匀量化

本文在 HSV 颜色空间下提取颜色特征，因为 HSV 颜色与人眼能感知的颜色一一对应，符合人眼的主观感觉。该模型对应于圆柱坐标系的一个圆锥形子集，圆锥的顶面对应于 $V=1$，代表的颜色较亮。色调 H 由绕 V 轴的旋转角给定，红、绿、蓝分别对应于角度 0°、120°、240°。每一种颜色和它的补色相差 180°。饱和度 S 的取值从 0 到 1，由圆心向圆周过渡 。

首先将图像中的 RGB 颜色转换为 HSV，然后进行颜色量化。颜色量化优化了不同颜色的使用，在保证图像视觉特性的同时，能够降低计算的复杂性。人眼对颜色的色调(H)非常敏感，色调独立于观察点，主要描述图像的颜色特征，所以要对其细量化。亮度(V)反映了图像的形状特征，不依赖于图像的彩色信息，所以在保证色调的情况下亮度分量不能太粗。根据 HSV 颜色空间的特性做以下量化：

$$H=\begin{cases}0, & h\in(315,20]\\ 1, & h\in(20,40]\\ 2, & h\in(40,70]\\ 3, & h\in(70,100]\\ 4, & h\in(100,130]\\ 5, & h\in(130,160]\\ 6, & h\in(160,190]\\ 7, & h\in(190,230]\\ 8, & h\in(230,270]\\ 9, & h\in(270,290]\\ 10, & h\in(290,300]\\ 11, & h\in(300,315]\end{cases},\quad V=\begin{cases}0, & v\in[0.1,0.2]\\ 1, & v\in(0.2,0.35]\\ 2, & v\in(0.35,0.5]\\ 3, & v\in(0.5,0.6]\\ 4, & v\in(0.6,0.85]\\ 5, & v\in(0.85,0.9]\end{cases}$$

$$S=\begin{cases}0, & s\in(0.15,0.55]\\ 1, & s\in(0.55,1.0]\end{cases}$$

其中，$v<0.1$ 的颜色归为黑色，$s<0.15$ 而且 $v>0.9$ 的颜色归入白色，量化后共有 $2+12\times6\times2=146$ 种颜色。最后将所有量化的颜色利用公式 $L=2+12H+6S+V$ 进行编码，黑色编码为 0，白色编码为 1，编码范围为 0－145。利用此方法，颜色量化保证了图像的整体颜色特性，失真小，如图 2 所示。

(a1) 原图　(a2) 量化后图像

(b1) 原图　(b2) 量化后图像

图 2

3.2 改进的颜色聚合向量(ICCV)

颜色直方图是在图像检索系统中使用最广泛的一种颜色特征[19]，颜色直方图特征提取简单，颜色丢失少，但缺点是没有表达颜色空间的分布信息，从而降低了检索精度。针对此缺点，Pass 等人提出了颜色聚合向量(Color coherence vectors)[20]，其是颜色直方图的一种演变，核心思想是将属于直方图每一个 bin 的像素分为两部分，如果 bin 内的某些像素所占据的连续区域的面积大于或等于给定的阈值 τ，则将区域内的像素作为聚合像素，否则为非聚合像素。颜色聚合向量比直方图能达到更好的检索效果。

为了能够更加准确地表达图像色彩的分布信息，考虑到图

像中聚合像素与连通区域、非聚合像素与非连通区域一一对应的关系，本文提出了改进的颜色聚合向量算法(ICCV)来提取图像的颜色特征。如果一幅图像的像素个数为 N，则阈值 τ 和 N 的关系一般为：$\frac{\tau}{N}=0.01$，从图像左上角开始，对每种颜色做 8-连通标记处理[21]，根据阈值找到颜色编码值 L_k 的连通域个数 $c(L_k)$以及非联通域个数 $n(L_k)$，求得每个编码值的聚合数目 α_k 以及非聚合数目 β_k，则每种颜色的平均面积聚合数目为$\frac{\alpha_k}{c(L_k)}$，平均面积非聚合数目为$\frac{\beta_k}{n(L_k)}$。然后对其进行归一化处理(见式(1)、式(2))，使其不受图像尺度变化的影响。

平均面积聚合数目：

$$A_c(L_k)=\frac{\alpha_k}{c(L_k)N} \tag{1}$$

平均面积非聚合数目：

$$A_n(L_k)=\frac{\beta_k}{n(L_k)N} \tag{2}$$

如果 $c(L_k)=0$ 或 $n(L_k)=0$，则 $A_c(L_k)=0$ 或 $A_n(L_k)=0$，图像的 ICCV 为：$(L_k,A_c(L_k),A_n(L_k))$，其中 $k=1,2,\cdots,146$，$L_k=0,1,2,\cdots,145$。如图 3 所示，如果黑色代表量化之后的编码值为 L_k 的像素，假设阈值为 3，那么其连通域个数 $c(L_k)$分别为 4、2、0，非联通域个数 $n(L_k)$分别为 0、2、6，聚合数目 α_k 分别为 14、7、0，非聚合数目 β_k 分别为 0、5、11，得到 $A_c(L_k)$分别为$\frac{14}{4N}$、$\frac{7}{2N}$、0，$A_n(L_k)$分别为 0、$\frac{5}{2N}$、$\frac{11}{6N}$，所以 ICCV 分别为$(L_k,\frac{14}{4N},0)$、$(L_k,\frac{14}{4N},\frac{5}{2N})$、$(L_k,0,\frac{11}{6N})$。

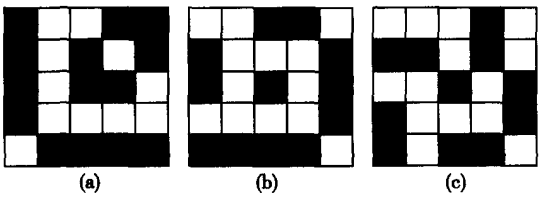

图 3

提取颜色特征是图像检索的关键一步，关系到图像检索的效果。以编码值 L_k 为横坐标，$x_{L_k}=(A_c(L_k),A_n(L_k))$为纵坐标，就构成了图像的颜色聚合向量直方图，可以用 146 维的向量 $\vec{X}=(x_{L_0},x_{L_1},\cdots,x_{L_{146}})$来表示一幅图像的特征，则可用 $N_p\times146$ 的矩阵来表示有 N_p 幅图像的特征数据库：

$$\begin{bmatrix}\vec{x}_1\\ \vec{x}_2\\ \vec{x}_3\\ \cdots\\ \vec{x}_N\end{bmatrix}=\begin{bmatrix}x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,146}\\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,146}\\ x_{3,1} & x_{3,2} & x_{3,3} & \cdots & x_{3,146}\\ & & \vdots & & \\ x_{N,1} & x_{N,2} & x_{N,3} & \cdots & x_{N,146}\end{bmatrix}$$

4 聚类

4.1 相关知识

在图像检索中，聚类的最终目的是建立索引特征库，减少对图像库的访问，提高检索的速度。本文提出一种基于贡献度的聚类算法将图像库中相似的图像聚类，建立特征索引库，在保证速度的同时提高检索效果。

贡献的概念起源于博弈论[22]，确定一个向量对其所在的类的贡献是对其所在类的质量的影响，根据这个度量标准，我们从图像特征库中选取最优的 K 个类。加格的一个研究[23]重点是博弈论和数据集的合并，其在合作博弈中将类的形成映射到联盟的形成中，并且利用夏普利值的解决方案找到给定点的最优的聚类数量。我们利用该方法并对其加以改进，核心思想是将特征向量划分到固定数量的类中。

$$I_a=\frac{1}{n_k}\sum_{\vec{x}_i\in\vec{c}_k}(\vec{x}_i-\vec{c}_k)^2 \tag{3}$$

其中，I_a 是类 $\vec{S}_k$ 的类内分散度；n_k 是类 $\vec{S}_k$ 包含的特征向量个数，即图像库中的图像；$\vec{c}_k$ 是类 $\vec{S}_k$ 的类心向量。一个向量 $\vec{x}\in\vec{c}_k$ 的贡献如式(4)所示：

$$Con(\vec{x}_i,\vec{S}_k)=I_a(\vec{S}_k-\{\vec{x}_i\})-I_a(\vec{S}_k) \tag{4}$$

可以容易地分析出：当一个向量的贡献是负的，其对所在的类有不利的影响；相反，如果贡献是正的，那么将这个向量从其所在的类中移除将会降低此类的质量(使其类内分散度变大)。除了考虑了由类内分散度引出的贡献度，在迭代优化时，还考虑了类间的分散度，分散度函数如下所示：

$$I_r=\frac{1}{K}\sum_{k=1}^{K}(\vec{c}_k-\vec{c})^2 \tag{5}$$

其中，I_r 是类间分散度，K 是类心向量的个数，$\vec{c}$是所有类心向量的平均值，要使聚类效果最好，就要让 I_a 足够小，I_r 足够大。所以，算法的主要思想是：如果一个向量的贡献是负的，我们将这个向量移动到另一个类中，使其贡献度最大；相反，如果一个向量的贡献是正的，那么我们使用优化规则综合类内和类间的分散度，并对它进行优化，使得聚类效果最好。

4.2 选取初始聚类中心

首先选择两个距离最大的初始聚类中心[12]，利用式(6)找到最大的 $D(\vec{x}_i,\vec{x}_j)$，则前两个类心为$\vec{c}_1=\vec{x}_i$；$\vec{c}_2=\vec{x}_j$。接下来计算之后的 $K-2$ 个类心，假设已经选择了 k 个类心，计算余下的特征向量与之前选择的 k 个类心向量的累加和，累加和最大的向量选为第 $k+1$ 个类心向量$\vec{c}_{k+1}$，公式如下：

$$D(\vec{x}_i,\vec{x}_j)=\sqrt{\sum_{k=1}^{146}(x_{i,k}-x_{j,k})^2} \tag{6}$$

$$x_{i,k}-x_{j,k}=|A_c^i(L_k)-A_c^j(L_k)|+|A_n^i(L_k)-A_n^j(L_k)| \tag{7}$$

$$S_i=\sum_{j=1}^{k}(\vec{x}_i-\vec{c}_j)^2 \tag{8}$$

其中，$x_{i,k}$表示第 $k(k\in(1,146))$个编码值对应的纵坐标分量$(A_c(L_{k-1}),A_n(L_{k-1}))$，两张图像的同一编码值之间的相似度使用绝对距离[20]比较好，在不影响结果的情况下可以减少计算量，而所有编码值的相似度采用欧氏距离来保证相似度的质量。找到最大的 $S_M=\max_i\{S_i\}$，则特征向量 $\vec{x}_M$ 就是第$k+1$个类心向量，依次类推，找到 K 个初始聚类中心。最后根据式(8)将每一个颜色特征向量分别分配到与其最接近的类心所在的类。

4.3 迭代优化

在 4.2 节已经得到了 K 个初始聚类中心，以及每个类具有的颜色特征向量，之后进行迭代优化，在此阶段，综合考虑了类内和类间的分散度，其中，$m[i]$是类心的索引，$I_{a,n}$和 $I_{r,n}$ 分别是将 $\vec{x}_i$ 加入到类 $\vec{S}_p$ 之后更新得到的 I_a 与 I_r 值。具体算法如算法 1 所示。

算法 1 本文聚类阶段算法

```
for 每一个类 S_k
  for 每一个特征向量 x_i
  if Con(x_i, S_k)≤0
    m[i]= argmax_{p∈{1,2,…,K}} Con(x_i, S_p)(其值为最大贡献度时 p 的值);
    将 x_i 加入到类 S_p(S_p={ x_i | m[i]=p} );
```

更新类心向量$\vec{c}_p$($\vec{c}_p=\frac{1}{n_p}\sum_{\vec{x}_j\in\vec{S}_p}\vec{x}_j$，$n_p$是类$\vec{S}_p$中特征向量的数目)；

end if

else if $Con(\vec{x}_i,\vec{S}_k)\geqslant 0$

将$\vec{x}_i$加入到类$\vec{S}_p$，使得$\frac{I_a-I_{a,n}}{I_a}+\frac{I_{r,n}-I_r}{I_{r,n}}$最大(比加入到其它类时的值大而且要比之前$\vec{x}_i$没有改变时的值要大)；

更新类心$\vec{c}_p$；

end

end

end

其中，当$Con(\vec{x}_i,\vec{S}_k)\geqslant 0$时，$\vec{x}_i$加入类$\vec{S}_p$的判定条件为$\frac{I_a-I_{a,n}}{I_a}+\frac{I_{r,n}-I_r}{I_{r,n}}$最大，有3种可能情况：第一种情况是$K$个类的类内分散度变大，即$\frac{I_a-I_{a,n}}{I_a}$变小，但是如果类间分散度变大，即$\frac{I_{r,n}-I_r}{I_{r,n}}$变大，且两者之和为最大，那么我们认为将$\vec{x}_i$加入到$\vec{S}_p$是可行的；第二种情况是如果将特征向量$\vec{x}_i$加入到类$\vec{S}_p$之后，类内分散度变小，即$\frac{I_a-I_{a,n}}{I_a}$变大，类间分散度变小，即$\frac{I_{r,n}-I_r}{I_{r,n}}$变小，但是两者之和最大，那也是可行的，因为这两种情况都是综合考虑了“类内相似度大，类间相似度小”这一准则；第三种条件是最理想的，即如果将特征向量$\vec{x}_i$加入到新类$\vec{S}_p$之后，类内分散度变小，即$\frac{I_a-I_{a,n}}{I_a}$变大，类间分散度变大，即$\frac{I_{r,n}-I_r}{I_{r,n}}$变大，两者之和最大，那么完全符合条件。

通过以上算法，我们可以得到最终的K个类$\{\vec{S}_1,\vec{S}_2,\cdots,\vec{S}_K\}$，以及类心$\{\vec{c}_1,\vec{c}_2,\cdots,\vec{c}_K\}$，在聚类结束之后，以每个聚类中心为索引项，建立图像特征的索引库，减少对图像库的访问次数，提高图像的整体检索速度。

5 查询图像

对于用户待查询图像I，提取其颜色直方图特征，得到特征向量$\vec{I}=(i_1,i_2,\cdots,i_D)$，利用式(9)计算$\vec{I}$到类心之间的距离，待查询图像属于与其距离最小的类。

$$D_{lk}=|\vec{I}-\vec{c}_k|=\sqrt{\sum_{m=1}^{146}(i_m-c_{k,m})^2} \tag{9}$$

其中，$1\leqslant k\leqslant K$，查询最小的$D_{lL}=\min_k|D_{lk}|$，则待查询图片属于$\vec{S}_L$类，再计算$\vec{S}_L$类中的每一个特征与$\vec{I}$的距离，距离越小相似度越大，按照相似度大小降序排列，返回$\vec{S}_L$中所有的图片。

6 实验验证

6.1 实验装置

软件的运行平台为Microsoft Visual Studio 2010，C++语言。在Corel图像库中选取2000幅图像(分为20类，每类有100幅)。在同样的计算机硬件环境下将本文算法与前面介绍的文献[8,11,12]算法进行比较。给定一幅图像，N为检索出的图像与目标图像相似的数目，T为检索出的图像总数目，R是库中图像的数目与目标图像相似的数目，采用查准率$Precision=\frac{T}{R}$和查全率$Recall=\frac{N}{R}$来表示检索的性能。

6.2 实验方案

在Corel图像库的20类图像中，每类抽取10幅图像进行检索，分别计算每一幅查询图像的查准率p_i以及查全率R_i，然后根据式(10)、式(11)计算每类图像的平均查准率$\bar{P}_{10}$以及查全率$\bar{R}_{10}$，其中，n为每类抽取的图像数，同理计算出$\bar{P}_{20}$与$\bar{R}_{20}$。按这种方法再做两次实验，每次使用的图像都不一样，计算总的平均查准率与查全率。

$$\bar{P}_n=\frac{1}{n}\sum_i^n P_i \tag{10}$$

$$\bar{R}_n=\frac{1}{n}\sum_i^n R_i \tag{11}$$

6.3 实验结果

参照以上方案对本文算法进行验证，其中图4和图5是图像检索的两个结果(左上角为待查询图像)。图4中有103幅图像，其中62幅是准确检索到的相关图像。图5中有108幅图像，其中61幅图像是准确检索到的相关图像。3种算法总平均查准率和总平均查全率的对比如图6和图7所示。结果很明显，本文算法的查准率和查全率均高于文献[8,11,12]3种算法。

图4 日落的检索结果

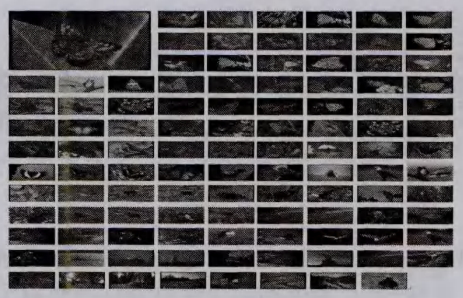

图5 蝴蝶的检索结果

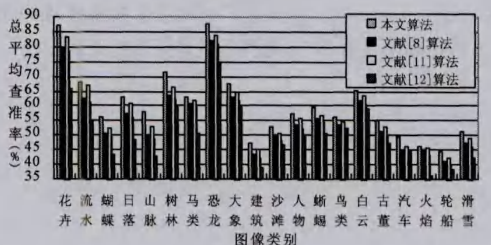

图6 各类图像总平均查准率比较

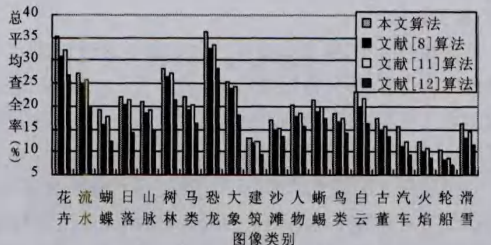

图7 各类图像总平均查全率比较

查全率-查准率曲线是衡量图像检索系统效果的一个重要指标，查准率和查全率存在互逆的关系，如图 8 所示，本文算法与文献[11]的查全率和查准率均比文献[8,12]高，而且本文算法曲线较平坦。质量好的图像检索系统其查准率-查全率曲线都比较平坦。因为本文检索系统建立了索引库，对同一图像库只需要一次特征提取，之后利用索引库可以重复使用，所以不考虑特征提取的时间，4 种算法的检索时间如表 1 所列，本文算法和文献[12]算法在 2000 幅图像中的检索速度较快且差别不大，而文献[8,11]算法的时间复杂度比较大。综合检索速度和查询效果来看，本文提出的算法更好。

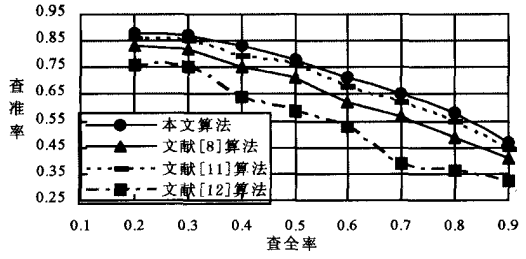

图 8　查准率-查全率曲线

表 1　4 种算法的图像检索时间

算法	图像检索时间
本文算法	10.6 s
文献[8]算法	13.5 s
文献[11]算法	20.2 s
文献[12]算法	9.8 s

结束语　本文提出了一种新的基于贡献度划分聚类的图像检索方法，利用改进的颜色聚合向量算法提取图像的颜色特征，利用累加和公式选取初始类心，避免了随机选取初始类心导致的聚类结果的不稳定性，引入了分散度和贡献度概念，优化了类内和类间的分散度，使聚类效果更好。实验表明，本文算法在保证较高检索速度的同时，具有更高的查准率和查全率。将各种视觉特征融合，并使用本文划分聚类算法进行检索，将是今后的研究重点。

参 考 文 献

[1] Dahane G M, Vishwakarma S. Content Based Image Retrieval System[J]. IJEIT, 2012, 1(5): 92-96

[2] 徐久成，任金玉，孙林，等. 基于云模型和相容粒的彩色图像检索方法[J]. 计算机科学，2013，4(12)：81-85

[3] Singhai N, Shandilya S K. A survey on: content based image retrieval systems[J]. International Journal of Computer Applications, 2010, 4(4): 22-26

[4] Chundi P, Dayal U, Sayal M, et al. A document clustering method and system[P]. US, 20077181678. 2007

[5] Jain M, Singh S K. A Survey On: Content Based Image Retrieval Systems Using Clustering Techniques For Large Data sets[J]. International Journal of Managing Information Technology, 2011, 3(4): 23-39

[6] Chen Y X, Wang J Z, Krovetz R. Content-based image retrieval byclustering[C]// Proceeding of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval. New York: ACM, 2003: 193-200

[7] Zhou H Y, Sadka A H, Swash M R, et al. Content-based image retrieval and clustering: a brief survey[J]. Recent Patents on Electrical Engineering, 2009, 2(3): 187-199

[8] Murthy V, Vamsidhar E, Kumar J S, et al. Content based image retrieval using Hierarchical and K-means clustering techniques [J]. International Journal of Engineering Science and Technology, 2010, 2(3): 209-212

[9] Muller K, Mika S, Ratsch G, et al. An introduction to Kernel-based learning algorithms [J]. IEEE Transactions on Neural Networks, 2001, 12(2): 181-201

[10] Górecki P, Sopyla K, Drozda P. Ranking by K-means voting algorithm for similar image retrieval[C]// Artificial Intelligence and Soft Computing. Springer Berlin Heidelberg, 2012: 509-517

[11] Ho J M, Lin S Y, Fann C W, et al. A novel content based image retrieval system using K-means with feature extraction[C]// 2012 International Conference on Systems and Informatics (ICSAI). IEEE, 2012: 785-790

[12] 吕明磊，刘冬梅，曾智勇. 一种改进的 K-means 聚类算法的图像检索算法[J]. 计算机科学，2013，40(8)：285-288

[13] Chang R I, Lin S Y, Ho J M, et al. A novel content-based image retrieval system using K-means/KNN with feature extraction [J]. Computer Science and Information Systems/ComSIS, 2012, 9(4): 1645-1661

[14] Lin C H, Chen C C, Lee H L, et al. Fast K-means algorithm based on a level histogram for image retrieval[J]. Expert Systems with Applications, 2014, 41(7): 3276-3283

[15] Havens T C, Bezdek J C, Leckie C, et al. Fuzzy c-means algorithms for very large data[J]. Fuzzy Systems, IEEE Transactions on, 2012, 20(6): 1130-1146

[16] Huang Z X. Extensions to the K-means algorithm for clustering large data sets with categorical values[J]. Data Mining Knowledge Discovery, 1998, 2(3): 283-304

[17] Chen Y X, Wang J Z, Krovetz R. CLUE: Cluster-based retrieval of images by unsupervised learning[J]. IEEE Transactions on Image Processing, 2005, 14(8): 1187-1201

[18] Zakariya S M, Ali R, Ahmad N. Combining visual features of an image at different precision value of unsupervised content-based image retrieval [C]// 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, 2010: 1-4

[19] 刘广海，吴璟莉. 基于颜色体积直方图的图像检索[J]. 计算机科学，2012，39(1)：273-275，280

[20] Pass G, Zabih R. Histogram refinement for content-based image retrieval[C]// Proceedings 3rd IEEE Workshop on Applications of Computer Vision, 1996(WACV'96). IEEE, 1996: 96-102

[21] Youngeun A, Junguk B, Sangwook S. Classification of featureset using k-means clustering from histogram [C]// Proceedings of the IEEE International Conference on Networked Computing and Advanced Information Management. 2008, 2: 320-324

[22] Osborne M J. An introduction to game theory[M]. USA, Oxford University Press, 2007: 41-43

[23] Garg V K. Pragmatic data mining: Novel paradigms for tackling key challenges [D]. Bangalore: Indian Institute of Science, 2009