

通过遗传聚类算法实现基于内容的图像检索

Implement of Image Retrieval Based on Genetic Algorithm to do the Clustering

陈 旌 (中国科学技术大学研究生院电子学部 100039)

摘要:本文介绍的算法使用了颜色直方图作为检索特征,利用遗传聚类算法对图像库的聚类结果,来实现基于内容的图像检索,此外,在算法中还融合了用户反馈技术来提高检索的准确率。

关键词:图像检索 颜色直方图 遗传聚类算法 用户反馈

1 引言

随着互联网的不断普及和数字化技术的快速发展,数字图像日益丰富,因此基于内容的图像检索(CBIR, Content Based Image Retrieval)已经成为当前的一个热门研究课题。

本文介绍的一种算法,就是以图像颜色的直方图作为图像的检索特征(即将图像间的相似程度归结为其颜色直方图间的差距),在检索过程中,首先使用遗传聚类算法根据图像间颜色直方图差距的大小,对图像库进行聚类,接着计算出目标图像归属于图像库中的哪个类别,最后将此类别中与目标图像距离最近的图像作为检索结果进行输出。由于考虑到计算机自动提取的图像特征与人所理解的语义之间存在巨大的差异,其检索结果往往难以令人满意,在此算法中,还融合了用户反馈技术:在检索过程中允许用户对检索结果进行评价和标记,指出结果中哪些与目标图像相似,哪些不相似,然后利用反馈信息修改图像特征向量的各分量的权值,突出其中较为重要的分量,来提高检索的准确率。

2 遗传聚类算法

对图像库进行聚类,就是把库中的图像按照其特征的相似程度划分成若干类别。设图像库中所有图像的特征向量集合为 $Z = \{Z_1, Z_2, \dots, Z_n\}$, Z_i 为图像库中第 i 幅图像的特征向量。此时,聚类就等价于将特征库中的特征向量根据其相互间距离(本文中利用的是欧式距离)的大小分为 k 类。通常使用的方法是为一类选一个图像作为凝聚点来代表这个类,并计算此

类中其他图像同凝聚点之间的距离,取其平均值作为衡量聚类结果的标准。凝聚点所组成的集合 R 是集合 Z 的一个子集。这样聚类问题就演变为一个组合优化问题,解这个问题的目标即为寻找一个由凝聚点所组成的集合使得如下公式取得最小值:

$$f(R) = \sum_{i=1}^n d(Z_i, \text{rep}[Z_i, R]) \quad (4)$$

其中 R 为 P 的子集,且 $|R| = k$ 。 d 代表图像特征向量间的距离。如果 $\text{rep}[Z_i, R]$ 是 $R = \{m_1, \dots, m_k\}$ 中距 Z_i 最近的图像,那么 $\min_{j \in \{1, \dots, k\}} d(Z_i, \text{rep}[Z_i, R])$ 。此时,搜索空间 $X = 2^P$, 其中 $R \subseteq Z$ 且 $|R| = k$ 。

由于图像数据库中的图像通常都是很多的,搜索空间 X 将会非常大,用穷举法进行搜索是根本不可行的,而传统的 K 均值聚类算法又有很容易陷入局部最优,因此本文将使用遗传算法对目标函数进行优化。具体计算步骤如下:

(1) 确定染色体的编码方法。本文使用的编码方法是将每一个可能的可行解编码为一个整数数组。其中,每一个染色体为一个含有 k 个整数的向量 C 。对于 $R \subseteq Z$, 且仅当 $\exists i | C[i] = i$ 时 $Z_i \in R$ 。例如,对 1000 个个体聚为 10 类,首先将它们编号为 1, 2, ..., 1000。一个可能的可行解数组 $X = \{980, 56, 10, 23, 653, 357, 228, 469, 300, 167\}$ 就表示一个个体的基因型,其中每一个基因座中的整数表示一个凝聚图像点的编号。

(2) 初始化。确定运行参数,主要有个体编码串长度 k (本文中取 256)、群体大小 M (本文中取 100)、重组概率 P_c (本文中取 0.66)、变异概率 P_m (本文中取 0.01)、终止代数 T (本文中取 200)。置遗传代数 $t = 0$,

随机产生 M 个初始个体,形成初始群体 $P(0)$ 。

(3) 计算群体 $P(0)$ 中各个个体的目标函数和适应度。在本文中,个体的目标函数为 $f(R) = \sum_{i=1}^n d(Z_i, \text{rep}[Z_i, R])$, 其中 R 中只有非负值参加 $\text{rep}[Z_i, R]$ 的计算, Z_i 表示特征库中的向量, d 代表欧氏距离, 即:

$$d(Z_x, Z_y) = \sqrt{\sum_{i=0}^m (z_{xi} - z_{yi})^2} \quad (5)$$

其中 Z_x, Z_y 分别表示经过 Z -score 变换后的图像特征向量, z_{xi}, z_{yi} 表示特征向量中的分量值, 由于本文使用的颜色直方图特征向量有 64 维分量, 因此 i 的取值为 0 到 63, 即 $m = 63$ 。本文寻求的目标是此公式的最小值, 因此必须对此目标函数值 $f(R)$ 到搜索空间中对应个体的适应度函数值 $F(R)$ 进行下述转换:

$$F(R) = \begin{cases} C_{\max} - f(R), & \text{if } f(R) < C_{\max} \\ 0, & \text{if } f(R) \geq C_{\max} \end{cases} \quad (6)$$

其中, C_{\max} 为一个适当地相对较大的数。

(4) $t = t + 1$

(5) 个体选择、复制操作。这里的选择算子使用的是赌盘选择, 个体被选中的概率取决于其适应度的大小。具体的执行过程是: 先计算出群体中所有个体的适应度的总和, 其次计算出每个个体的相对适应度的大小 $F_i(R) / \sum_{i=1}^M F_i(R)$ (其中 i, j 表示群体中的个体编号, M 表示群体的总数), 它即为各个个体被遗传到下一代群体中的概率。最后再使用模拟赌盘操作 (即 0 到 1 之间的随机数) 来确定各个个体被选中的个数。

(6) 个体重组操作

(7) 个体变异操作。这里使用的变异算子是本位变异算子, 即以很低的概率 P_m 将集合 R 中的元素 $Z_i \in R$ 替换为 $Z_i \notin R$ 。

(8) 评价群体的适应度。

(9) 终止条件判断。若不满足终止条件, 则转到第 4 步, 继续进化过程; 若满足终止条件, 则输出进化过程中的最优个体, 并除去此个体中所有值为负的基因座得到实际的可行解数组, 算法结束。

3 图像检索算法

在本文中, 基于内容的图像检索步骤如下:

(1) 提取图像库中被检索图像的颜色直方图特征向量。组成由式 (3) 表示的特征库。

(2) 使用 2.2 中介绍的遗传聚类算法对特征库进行聚类。计算目标图像与各个类别凝聚图像点间的距离, 判断目标图像属于哪个类别 (与此类别的凝聚图像点间的距离最小)。

(3) 将此类别中的与目标图像距离最近的 l 幅图像反馈给用户。如果此类别中包含的图像数少于 l 幅, 可以选取其它类别中与目标图像距离最近的补足剩余的图像。

(4) 用户对反馈的图像进行评价。与目标图像相似的图像给予正向评价, 设其数目为 l_p , 不相似的图像给予负向评价, 设其数目为 l_n , 其中 $l = l_p + l_n$ 。

(5) 将所有被给予正向评价的图像与目标图像组合在一起计算它们特征向量的每一维分量的标准方差, 分别为 S_0, S_1, \dots, S_{63} , 并以这些标准方差的倒数的开方 ($w_i = 1/\sqrt{S_i}$) 作为相应特征分量的权值, 分别为 w_0, w_1, \dots, w_{63} 。用这些权值分别乘以图像的原始特征向量的相应分量, 得到图像的新特征向量 $Z_i (w_0 z_{i,0}, w_1 z_{i,1}, \dots, w_{63} z_{i,63})$ 。对图像库中的所有图像特征向量完成此项操作后, 剔除掉被给予负向评价的图像构成新的特征库。

(6) 转到步骤 2 对新的特征库进行聚类, 并继续进行检索, 直到找到足够数量的相似图像, 或者相似图像数目不再增加为止 (一般需要进行 5 到 10 次)。

4 实现与实验结果

为了验证上面的算法, 本文利用 Matlab 6.5 实现了一个试验系统, 该系统带有一个测试图像数据库, 库中共有 6000 幅图像, 图像的分辨率 384×256 或者 256×384 , 它们是从 "COREL" 中挑选出来的, 一共 40 个分类, 每个分类包含的图像数目从 10 到 200 不等。实验中从每一个类别中随机选出 3 幅图像作为目标图像, 这样总共检索 120 幅目标图像。对每一幅目标图像, 用户需要进行 10 次检索, 每次检索系统返回 18 (即 $l = 18$) 幅图像。试验过程为: 用户根据初次检索的结果, 选择相似的图像作为反馈, 系统根据反馈计算权值, 并调整特征库中的各个特征向量, 形成新的特征库, 以新的特征库为基础, 继续作二次检索, 如此循环 10 次, 图 1 和图 2 展示了其中一幅图像

的检索结果。

于 18, 则每次检索的准确率为 $precision = I_n/18$, 如果小

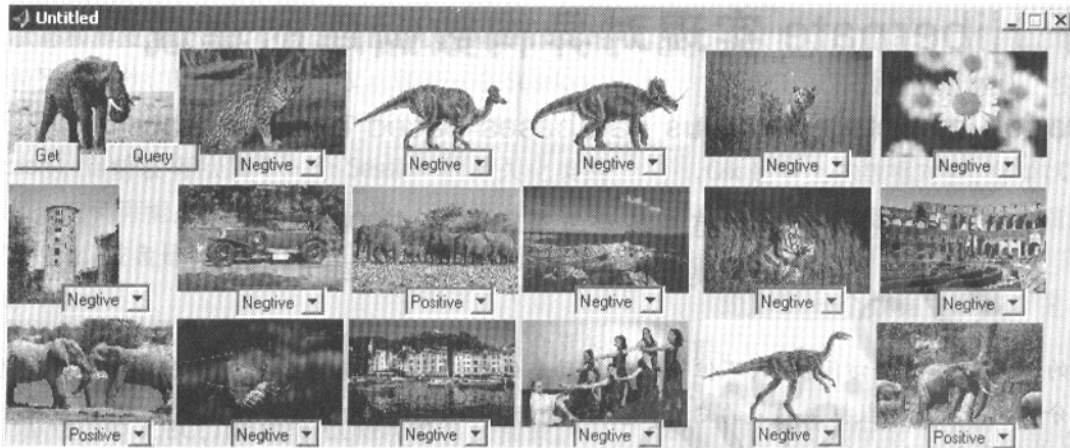


图 1 首次检索结果

于 18, 则每次检索的准确率为 $precision = I_n/q$ (q 为此类别中包含的图像数)。综合整个试验, 得到了相应的平均准确率, 如图 3 所示。从实验中可以看出, 经过多次检索, 检索准确率达到 0.8。

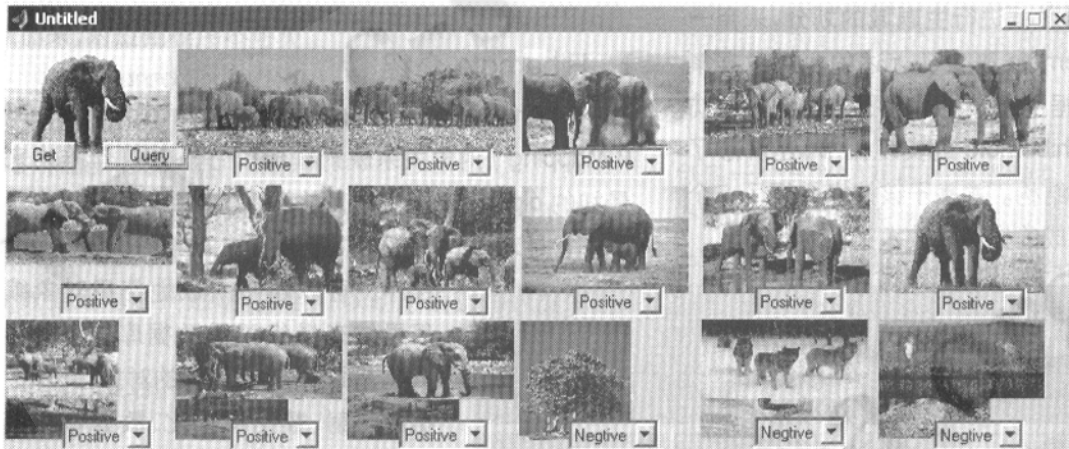


图 2 第 10 次检索结果

5 结束语

本文介绍的算法以图像颜色的直方图作为图像的检索特征, 利用遗传聚类算法对图像库的聚类结果, 来实现基于内容的图像检索, 此外, 在算法中还融合了用户反馈技术。实验结果

表明, 这种算法是可行的, 且具有简洁和高效的特点。实际上, 用于检索的图像特征包括很多种, 如纹理、灰度、形状等也很重要, 也可以应用本文的思想使用这些特征来进行图像检索。

参考文献

- 1 田玉敏、林高全, 基于颜色特征的彩色图像检索方法 [J], 西安电子科技大学学报(自然科学版), 2002, 29(1)。
- 2 Swain MJ, Ballard DH. Color indexing [J]. International Journal of Computer Vision, 1991, 7(1)。
- 3 陈国良等, 遗传算法及其应用 [M], 人民邮电出版社, 1996 Ali K, et al.

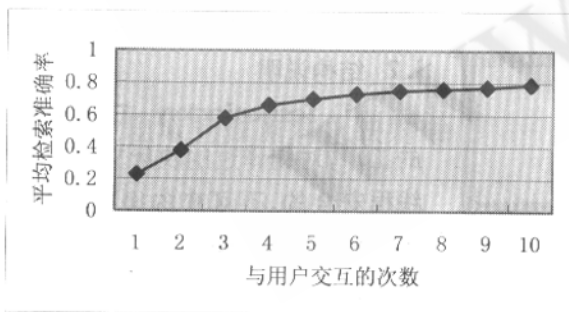


图 3 检索的准确率

实验中规定, 判断两幅图像是否相似的依据是它们是否属于同一类别, 如果此类别中包含的图像数大