

新的改进 K-means 算法的图像检索方法

马娟娟¹, 张 伟¹, 李朝锋¹, 杨 弘²

(1 江南大学 物联网工程学院, 江苏 无锡 214122;

2 中国水产科学研究院 淡水渔业研究中心, 江苏 无锡 214081)

摘 要: 提出一种改进 K-means 算法的初始类心选取方法. 首先基于 HSV 颜色空间计算样本图像和图像库中所有图像的特征向量, 将样本图像的特征向量作为第一个初始类心, 然后计算图像库所有图像距离样本图像的特征向量距离, 找出距离最大的特征向量作为第二个初始类心, 在剩下的特征向量中找到距离前两个初始类心和最大的特征向量作为第三个初始类心, 依次类推确定剩下的初始类心, 然后进行聚类, 最后进行图像检索. 实验结果论证了此算法的有效性.

关键词: 图像检索; K-means 算法; HSV 颜色空间; 初始类心; 特征向量

中图分类号: TP391.413

文献标识码: A

文章编号: 1000-7180(2014)12-0148-04

An Image Retrieval Method Based on a New Improved K-means Algorithm

MA Juan-juan¹, ZHANG Wei¹, LI Chao-feng¹, YANG Hong²

(1 School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China;

2 Freshwater Fisheries Research Center, Chinese Academy of Fishery Science, Wuxi 214081, China)

Abstract: According to the shortcomings of instability and low efficiency for K-means based image retrieval algorithm, this paper proposed an improved method to determine the initial class center of K-means clustering algorithm. Firstly, it uses HSV color space to calculate the feature vectors of sample image and all the images in the image database, and take the sample image's feature vector as the first initial cluster center, then calculate the distances of feature vector between all the images in the image database and sample image, and take the feature vector which has biggest distance as the second initial cluster center, and take the remaining feature vector of the most far away from the two initial cluster centers as the third initial cluster center, and determine the other initial cluster centers in the same way. Finally, the clustering are done according to the initial cluster centers to finish image retrieval. Experimental results suggest the validity of the algorithm.

Key words: image retrieval; K-means algorithm; HSV color space; initial cluster center; feature vector

1 引言

基于内容的图像检索(CBIR)技术^[1]已经成为研究的焦点. CBIR 技术主要根据图像的颜色、纹理、形状以及对象的空间关系等底层图像特征来进行图像间的匹配^[2],进而快速地搜索相关图像信息.

但是直接利用颜色特征进行图像检索效率不高,如果能够依据图像之间的排斥性和吸引力对图像库进行有效分类^[3],就能立刻缩小检索的范围和提高图像检索的速度.

K-means 算法在图像检索中应用有很多^[4],它是依据初始类心把各自相似的数据项依据最小距离

收稿日期: 2014-02-11; **修回日期:** 2014-04-07

基金项目: 国家自然科学基金项目(61170120);教育部新世纪人才计划(NCET-12-0881);现代农业产业技术体系专项资金资助(CARS-49)

归并到不同的类. K-means 算法的初始类心是随机产生的,所以算法结果存在不稳定性. 文献[4]中计算图像库所有图像基于 RGB 颜色空间的特征向量,其中使距离最大的两个特征向量作为前两个初始类心,在剩下的特征向量中找到与前两个初始类心距离之和最大的特征向量作为下一个初始类心,按照此方法直到查找到 k 个初始类心,获得了稳定的图像检索结果. 依据文献[4]中的算法计算初始类心,图像库是固定不变的,不同样本图片检索时初始类心都是相同的,进而图像库的聚类过程是相同的. 为了使图像库能够依据不同样本图像进行相应的聚类,本文对初始类心选取方法作了改进. 实验结果表明了本文方法的有效性.

2 提取图像的颜色特征

颜色是现有检索系统中应用最广泛的特征. 常用的颜色空间模型有很多种. 数字图像一般采用 RGB 空间模型表示,RGB 颜色模型是工业界的一种颜色标准,是从颜色发光的定义的一种模式,可分辨色差是非线性的. HSV 颜色模型是基于人的眼睛对色彩的识别^[5],是一种从视觉的角度定义的颜色模型. H 表示为色调, S 表示为饱和度, V 表示亮度,通过调整色调,饱和度及亮度得到颜色和变化,因此本文采用 HSV 颜色空间模型.

由于图像库读取的图像都是 RGB 颜色空间模型的,首先将 RGB 颜色空间模型转化为 HSV 颜色空间模型,转换模型关系如下:

$$h = \begin{cases} \arccos\left\{\frac{(R-G)+(R-B)}{2\sqrt{(R-G)^2+4(R-B)(G-B)}}\right\}, & B \leq G \\ 2\pi - \arccos\left\{\frac{(R-G)+(R-B)}{2\sqrt{(R-G)^2+4(R-B)(G-B)}}\right\}, & B > G \end{cases} \quad (1)$$

$$s = \frac{\max(R, G, B) - \min(R, G, B)}{\max(R, G, B)} \quad (2)$$

$$v = \frac{\max(R, G, B)}{255} \quad (3)$$

设 RGB 颜色空间值为 (R, G, B) , 其中 $R, G, B \in [0 \sim 255]$, 由式(1)~(3)分别得到 h, s, v 的值, 其中 $h \in [0 \sim 360]$, $s \in [0 \sim 1]$, $v \in [0 \sim 1]$. 依据色彩的不同范围将 h, s, v 分量进行非等间隔的量化,取色调 h 分成 8 份,饱和度 s 分成 3 份,亮度 v 分成 3 份,将量化后的各分合成为一维特征矢量,即:

$$L = HQ_sQ_v + SQ_v + V \quad (4)$$

式中, Q_s, Q_v 分别是分量的量化级数,通常取 $Q_s = 3, Q_v = 3$, 这样减少图像亮度和饱和度对图像色调的影响,因此式(4) 可以表示为

$$L = 9H + 3S + V \quad (5)$$

其中色调的权值为 9, 饱和度的权值为 3, 亮度的权值为 1, 然后计算 L 的一维直方图表示为

$$I(i) = \frac{n_i}{N} (i = 0, 1, \dots, L-1) \quad (6)$$

式中, L 表示一维量化后颜色种类, n_i 是图像中颜色值为 i 的像素个数, N 为图像像素的总数. 由于式(6) 中直方图常出现一些零值, 这些零值会对相似度计算带来影响^[6], 故本文采用累加直方图法即:

$$I(m) = \sum_{i=0}^m \frac{n_i}{N} (m = 0, 1, \dots, L-1) \quad (7)$$

$I(m)$ 表示图像中颜色值为 0 到 m 的像素占整个图像像素的比例和. 因此任意一张图像的特征向量可以表示为一个多维的向量.

3 改进 K-means 算法

K-means 算法是硬聚类算法, 是典型的基于距离的聚类算法, 采用欧式距离作为相似性的评价指标, 即认为两个对象的距离越近, 其相似度就越大. 具体流程^[3] 如下:

(1) 从 n 个数据对象任意选择 k 个对象作为初始类心;

(2) 计算每个对象与类心的距离, 并根据最小距离归并到相应的类;

(3) 重新计算各自新产生类的均值, 同时使它们分别成为各自类新的类心;

(4) 循环(2) 到(3) 直到每个聚类不再发生变化为止.

3.1 初始类心的改进

K-means 算法中初始类心是随机产生的, 即任意选取 k 个对象作为初始类心. K-means 算法中同一张图像不同次检索的初始类心是不同的, 文献[4] 中对初始类心做了改进, 获得了稳定的检索结果. 但是不同样本图像检索时使用同一个初始类心, 导致不同样本图像进行检索的聚类过程和结果都是相同的, 同时初始类心的选取缺少了样本图像的信息, 因此本文对初始类心做了改进.

首先将所有图像在 HSV 颜色空间下进行量化, 计算图像库所有图像的颜色特征向量, 假设图像库有 n 张图像, 由于每张图像都不一样, 所以有 n 个特征向量, 特征向量集为 $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots, \mathbf{X}_n)$, 每

个特征向量有 m 维,任意特征向量 $\mathbf{X}_i$ 与特征向量 $\mathbf{X}_j$ 之间的欧式距离即:

$$|\mathbf{X}_i - \mathbf{X}_j| = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + \cdots + (X_{im} - X_{jm})^2} \quad (8)$$

设初始类心向量集为 $\mathbf{V}$, 即 $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \cdots, \mathbf{V}_k)$, 计算样本图像特征向量 $\mathbf{V}_{\text{sample}}$ 作为初始类心 $\mathbf{V}_1$, 依据式(8) 分别计算特征向量集 $\mathbf{X}$ 里所有特征向量与 $\mathbf{V}_1$ 的距离, 找出距离 $\mathbf{V}_1$ 最大的特征向量为第二个初始类心 $\mathbf{V}_2$, 接下来确定剩下的 $(k-2)$ 个初始类心. 假设要确定第 $s(2 \leq s \leq k)$ 个初始类心 $\mathbf{V}_s$, 则已经确定了 $(s-1)$ 个初始类心, 分别计算特征向量集 $\mathbf{X}$ 剩下的 $(n-(s-1))$ 个特征向量与已经确定的 $(s-1)$ 个初始类心距离之和, 计算第 i 个向量与 $(s-1)$ 个初始类心距离之和表示为

$$D_i = \sum_{j=1}^{s-1} |\mathbf{X}_i - \mathbf{V}_j|; 1 \leq i \leq n+1-s \quad (9)$$

$$|\mathbf{X}_i - \mathbf{V}_j| = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + \cdots + (X_{im} - X_{jm})^2} \quad (10)$$

在 D_i 中找出最大的距离 $D_{\max}$, 则与之对应的 $\mathbf{X}_{\max}$ 就作为第 s 个初始类心, 即 $\mathbf{V}_s = \mathbf{X}_{\max}$. 重复这样的步骤得到 k 个初始类心.

3.2 评价聚类性能的准则函数

通过上面初始类心的改进获取了 k 个初始类心, 接下来就是对特征向量集 $\mathbf{X}$ 依据 K-means 聚类流程(1) ~ (4) 进行聚类, 生成聚类为 $\mathbf{C}$, 即 $\mathbf{C} = (\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3, \cdots, \mathbf{C}_k)$, 聚类后的类心为 $\mathbf{V}$, 即 $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \cdots, \mathbf{V}_k)$, K-means 算法使用误差平方和准则函数来评价聚类性能, 误差平方和准则函数公式为

$$E = \sum_{j=1}^K \sum_{Y \in C_j} |\mathbf{Y} - \mathbf{V}_j|^2 \quad (11)$$

式中, E 表示迭代时所有对象的平方误差总和, 迭代次数越大, E 越小, 直到 E 收敛到某一个固定的值, 则聚类结束.

4 图像相似度计算

基于内容的图像检索的主要特点是样本图像与图像库图像之间的相似性匹配, 而不是精确匹配. 本文采用欧式距离来度量图像的相似性, 欧式距离越小, 图像越相似. 由于本文采用颜色特征描述图像, 因此两个图像相似度就是两个图像的颜色特征的欧式距离.

对图像库的图像特征向量集聚类后, 获得聚类后的类心 $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \cdots, \mathbf{V}_k)$, 生成聚类为 $\mathbf{C} = (\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3, \cdots, \mathbf{C}_k)$, 分别计算样本图像颜色特征向量 $\mathbf{V}_{\text{sample}}$ 与每个类心的欧式距离, 依据相似性计算原理, 距离越小, 相似度越大, 图像越类似, 距离最小的类心 $\mathbf{V}_{\min}$ 所在的类 $\mathbf{C}_{\min}$ 则是聚类 $\mathbf{C}$ 中与样本图像最相似的类, 并将 $\mathbf{C}_{\min}$ 代表的图像类从图像库分离出来, 然后分别计算样本图像特征向量 $\mathbf{V}_{\text{sample}}$ 与 $\mathbf{C}_{\min}$ 中每一个特征向量的欧式距离, 然后依据距离进行升序排序, 按照排序后输出图像.

5 实验结果

为了验证本文算法的有效性, 采用 Corel 图像库 1 000 幅图像进行测试, 有 10 个图像类别, 包括马群、恐龙、花卉、汽车、非洲等等, 每类图像为 100 幅. 从每类图像中抽取 10 幅不同的图像分别作为样本图像进行检索. 评价图像的检索性能采用查准率^[6] p 来衡量, 公式如下:

$$p = \frac{n}{N} \quad (12)$$

式中, n 代表检索出相关的图像数, N 代表检索返回的图像总数. 一张恐龙样本图像检索返回的图像总数为 100, 其中相关的图像数为 83, 则查准率为 0.83. 一张马群样本图像检索返回的图像总数为 116, 其中相关的图像数为 96, 则查准率为 0.83.

令 $k=10$, 分别计算 10 个图像类别的 3 种算法平均查准率.

由表 1 得知, 每类图像文献[4]的查准率都比 K-means 算法查准率高, 而本文算法在初始类心加入样本图像信息, 不同样本图像检索时初始类心都是不一样的, 能够根据不同样本图像对图像库进行相应的聚类, 故提高了查准率.

表 1 三种算法的平均查准率

图像类	K-means	文献[4]	本文
人脸	0.49	0.51	0.55
沙滩	0.38	0.42	0.46
建筑	0.30	0.32	0.35
汽车	0.32	0.37	0.42
花朵	0.63	0.68	0.79
马群	0.65	0.68	0.77
雪山	0.39	0.41	0.43
大象	0.30	0.32	0.36
恐龙	0.70	0.76	0.84
食物	0.19	0.20	0.23

将本文算法与基于累加直方图的检索分别进行

实验. 每类图像本文的查准率均高出基于累加直方图的检索(CH)许多. 本文在相似度计算前缩小了检索的范围,减少了与样本图像特征差异较大的图像产生的干扰,因而能够在查准率上提高了很多.

6 结束语

在基于内容的图像检索中,一般是直接用样本图像的特征与图像库图像的特征进行相似性度计算,忽略了图像库不同类别图像之间的排斥性,相似图像之间的吸引力. 本文通过改进 K-means 算法的初始类心选取方法对图像库进行聚类,提高了图像检索效率. 由于颜色特征作为单一的特征,不能够完全准确的描述图像内容,如何用其他特征描述图像并进行检索,是下一步工作,以进一步提高图像检索的准确率.

参考文献:

- [1] Velpkamp R C. Content-based image retrieval system: a survey[R]. Utrecht: Utrecht University, Mirela Tanase Department of Computing Science, 2001.
 - [2] Rui Y, Huang TS, Chang S-F. Image retrieval: current techniques, promising directions and open issues[J]. Journal of Visual Communication and Image Representation, 1999, 10(3): 39-62.
 - [3] Huang Yuan-yuan, He Yun-feng. An image retrieval method based on color features[J]. Journal of Chinese Computer Systems, 2007, 28(7) : 1277-1281.
 - [4] 吕明磊, 刘冬梅, 曾智勇. 基于改进 K-means 算法的图像检索方法[J]. 计算机应用, 2013, 33(S1): 195-198.
 - [5] 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): 127-131.
 - [6] 宋琳琳, 王相海. 一种基于 HVS 加权颜色特征的图像检索算法[J]. 微电子学与计算机, 2008, 25(6): 60-62.
 - [7] 张鑫, 温显斌, 孟庆霞. 基于颜色特征的图像检索方法研究[J]. 计算机科学, 2012, 39(11): 243-245.
 - [8] 常哲, 侯榆青, 李明俐, 等. 综合颜色和纹理特征的图像检索[J]. 微型计算机系统, 2011, 32(1): 161-164.
- 作者简介:**
马娟娟 女, (1990-), 硕士研究生. 研究方向为图像检索、软件工程. E-mail: majuanjuan129@163. com.
张伟 男, (1956-), 副教授. 研究方向为模式识别、软件工程.
李朝锋 男, (1971-), 教授. 研究方向为图像及视频质量评价, 人工智能, 模式识别.
杨弘 男, (1966-), 研究员. 研究方向为渔业经济与信息技术.
-
- (上接第 147 页)
- [12] 黄红星, 景林, 黄习培. 挖掘最大频繁项集的朴素蚁群优化算法[J]. 计算机工程与设计, 2011, 32(1): 301-304.
 - [13] Frequent Itemset Mining Implementations Repository [EB/OL]. [2014-03-23]. <http://fimi.cs.helsinki.fi/>, 2004.
- 作者简介:**
黄红星 男, (1979-), 硕士, 讲师. 研究方向为优化算法及其应用. E-mail: qingtiandaxia@126. com.