

基于分块加权颜色直方图的图像聚类算法研究

杨传慧,吉根林,章志刚

(南京师范大学计算机科学与技术学院,江苏 南京 210023)

[摘要] 提出采用分块加权颜色直方图作为图像特征,分别利用 Affinity Propagation (AP) 算法和 k -means 算法对彩色图像进行聚类,将两种图像聚类算法加以实现,进行算法性能研究. 实验结果表明,应用 AP 算法对图像聚类的效果优于 k -means 算法对图像聚类的效果.

[关键词] 图像聚类,颜色直方图,AP 算法, k -means 算法

[中图分类号] TP312 [文献标志码] A [文章编号] 1672-1292(2013)01-0040-05

Research on Image Clustering Algorithm Based on Block Weighted Color Histogram

Yang Chuanhui, Ji Genlin, Zhang Zhigang

(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China)

Abstract: The block weighted color histogram is introduced as the feature of the color image. Both affinity propagation (AP) algorithm and k -means algorithm are implemented separately for clustering images, and the efficiencies of the two algorithms are studied. The experiment results show that the precision of AP algorithm for clustering image is better than k -means algorithm.

Key words: image clustering, color histogram, AP algorithm, k -means algorithm

随着数字化与互联网技术的飞速发展以及电子数码设备的普及,人们能够获取的数字图像信息越来越多. 对于海量的图像数据,如何有效地进行管理检索,并从中获取潜在的有用信息,已成为人们关注的问题. 基于内容的图像检索和图像数据挖掘近年来也成为学者们研究的热点课题.

图像聚类是通过无监督学习方法将目标图像数据集分成若干类,使得同一类别的任意两幅图像具有较高的相似度,不同类别的图像具有较高的相异度. 在图像管理与检索中,图像聚类可对图像进行聚类预处理以提高检索性能,并可帮助用户发现其感兴趣的信息等. 目前,基于内容的图像聚类研究已取得较多的成果^[1-4].

由于颜色是图像最基本最接近人类视觉感知的特征,因此本文基于 HSV 颜色空间提取图像的颜色特征,采用分块加权颜色直方图作为图像特征,分别运用 Affinity Propagation (AP) 算法^[5]和 k -means 算法^[6]对彩色图像进行聚类,并对两种算法的性能进行了比较. 实验结果表明,应用 AP 算法对图像聚类的效果优于 k -means 算法对图像聚类的效果.

1 基于分块加权颜色直方图的图像特征提取

1.1 HSV 颜色空间非均匀量化

图像特征提取是进行图像聚类的基础,好的图像特征能够比较精确地标识一幅图像,为高效的图像聚类提供有力的数据支持. 图像的底层特征主要包括颜色、纹理和形状等,对应的特征提取方法主要有颜色直方图方法、Tamura 纹理方法、形状无关矩方法以及这些方法的派生^[7]. 颜色直方图的核心思想是在一定

收稿日期:2012-09-01.
基金项目:国家自然科学基金(40871176).
通讯联系人:吉根林,博士,教授,博士生导师,研究方向:数据挖掘技术及应用. E-mail: glji@njnu.edu.cn.

的颜色空间对图像各种颜色出现的频数进行统计,其主要优点是对图像的尺寸、位置和方向的依赖性较小,特征提取相对简单. HSV 颜色空间则是用色调(H)、饱和度(S)和亮度(V)来表示颜色,这种模式较好地反映了人类视觉系统对色彩的理解方式. 因此,本文基于 HSV 颜色空间计算图像的颜色直方图作为图像的特征.

一般图像的存储模式为 RGB 模式,故需先将 RGB 模式转换为 HSV 空间模型. 转换后,HSV 各分量的取值范围分别为: $h \in [0, 360]$, $s \in [0, 1]$, $v \in [0, 1]$. 一幅真彩色图像的颜色有很多种,实际处理中为了提高效率、降低计算复杂度,往往只选取少量有代表性的颜色. 为此,可对 HSV 颜色空间进行非均匀量化,具体量化方法如式(1)所示.

色调 H 从 0° 渐变到 360° , 不等间隔地分为 7 份, 分别对应人类所能感知的心理颜色: 赤、橙、黄、绿、青、蓝、紫. 对于亮度 V 和饱和度 S 则分别划分为 3 个等级.

$$H = \begin{cases} 0 & \text{if } h \in (330, 360] \text{ or } h \in [0, 22] \\ 1 & \text{if } h \in (22, 45] \\ 2 & \text{if } h \in (45, 75] \\ 3 & \text{if } h \in (75, 155] \\ 4 & \text{if } h \in (155, 186] \\ 5 & \text{if } h \in (186, 287] \\ 6 & \text{if } h \in (287, 330] \end{cases} \quad S = \begin{cases} 0 & \text{if } s \in [0, 0.2] \\ 1 & \text{if } s \in (0.2, 0.7] \\ 2 & \text{if } s \in (0.7, 1] \end{cases} \quad V = \begin{cases} 0 & \text{if } v \in [0, 0.2] \\ 1 & \text{if } v \in (0.2, 0.7] \\ 2 & \text{if } v \in (0.7, 1] \end{cases} \quad (1)$$

根据 H 、 S 、 V 的量化级数和其频带宽度,将这 3 个颜色分量按式(2)进行组合:

$$L = HQ_s + SQ_v + V, \quad (2)$$

其中, Q_s 、 Q_v 分别是 S 和 V 的量化级数,即 $Q_s = 3$, $Q_v = 3$, 则式(2)可以表示为:

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

L 的取值为 $0, 1, 2, \dots, 62$, 即一幅彩色图像中的每个像素通过上述量化方法映射到了 63 种颜色空间中.

1.2 分块加权颜色直方图

根据上节所介绍的方法对一幅图像进行颜色量化处理之后,即可计算该图像的颜色直方图. 这样直接计算出来的全局直方图,虽简单易行,但其忽略了颜色的空间分布关系,而是将图像中每个部位的颜色都看做同等重要. 事实上,一幅图像中不同部位的颜色提供的信息量是不同的,通常一幅图像的信息主要集中在图像的正中央,而边缘部分往往作为背景,因此可对图像进行简单的分块处理,每一块赋予不同的权重. 图像分块方法有等距离环形分块方法、不均匀分块方法等. 本文采用如图 1 所示的不均匀分块方法.

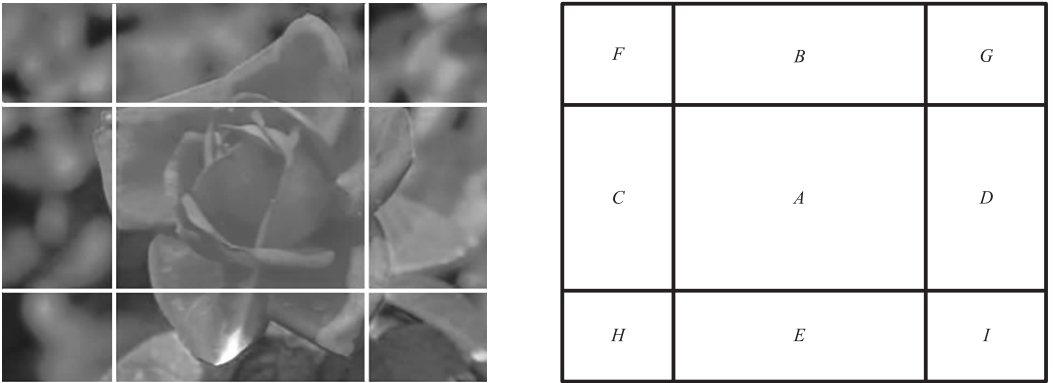


图 1 图像分块

Fig. 1 Image Blocking

图 1 中, A 区域位于图像中心, 包含了图像的主体信息, 赋予较大的权重; B 、 C 、 D 、 E 所包含的图像信息较少, 则赋予较小的权重; F 、 G 、 H 、 I 包含的图像信息最少, 赋予最小的权重. 所有权重的累加和为 1. 图像分块后, 按照上节方法分别计算每一块的颜色直方图, 然后各自乘上对应的权重, 最后累加得到整幅图像的分块加权颜色直方图, 即:

$$H(I)=\sum_{k=1}^9w_kH(I_k),\quad\sum_{k=1}^9w_k=1.$$

(4)

其中, I_k 表示图像的第 i 个分块.

2 图像聚类算法

传统的聚类方法如 k -means 等需要预先指定聚类数目,且对于初始聚类中心的选择非常敏感,当数据量较少且初始信息选择恰当时,可获得较好的聚类效果. 图像数据的特点是数据量大、种类繁多,因此传统的聚类算法不能很好地应用于图像聚类中. AP 算法是 Frey 等人 2007 年提出的一种新的无监督聚类算法^[5],文献[8]对该算法又进行了改进. AP 算法不需要预先指定聚类数目,聚类质量好、效率高,运行结果比较稳定,特别是在处理大数据量时效果尤为明显. 基于此,本文选用 AP 算法进行图像聚类,并与 k -means 聚类算法进行性能比较实验.

2.1 相似性度量方法

图像的相似性度量方法有多种,如欧式距离、二次式距离、卡方统计矩等. 本文通过实验研究表明,AP 算法选用线性核 (Linear kernel) 作为相似性度量方法,对于图像聚类能够取得较好的效果. 设两幅图像 X , Y 的特征向量分别为 $\boldsymbol{x}=(x_1,x_2,\cdots,x_d)$, $\boldsymbol{y}=(y_1,y_2,\cdots,y_d)$, 则 X 与 Y 的相似度为:

$$s(X,Y)=\frac{\sum_{i=1}^dx_iy_i}{\sqrt{(\sum_{i=1}^dx_i^2)(\sum_{i=1}^dy_i^2)}}.$$

(5)

2.2 AP 算法思想

AP 算法的主要思想是初始时将所有的数据点都看做候选聚类中心,在不断的迭代过程中,每个数据点通过相互之间“消息传递”来竞争聚类中心或者选择某个数据点作为自己的中心,最终获得若干个高质量的聚类中心. 这种消息传递机制避免了传统划分算法中随机选择模式对整个聚类结果的影响.

AP 算法的消息传递机制主要包含两种信息:吸引度 (Responsibility) 和归属感 (Availability). 吸引度 $r(i,k)$ 表示从数据点 i 发送到候选聚类中心 k 的消息,反映了 k 点作为 i 点聚类中心的合适程度;归属感 $a(i,k)$ 表示从候选聚类中心 k 发送到 i 的消息,反映了 i 点选择 k 点作为其聚类中心的合适程度. $r(i,k)$ 和 $a(i,k)$ 越大,则 k 点作为聚类中心的可能性就越大,且 i 点隶属于以 k 点为聚类中心的类别的可能性也越大. 图 2 描述了消息传递过程.

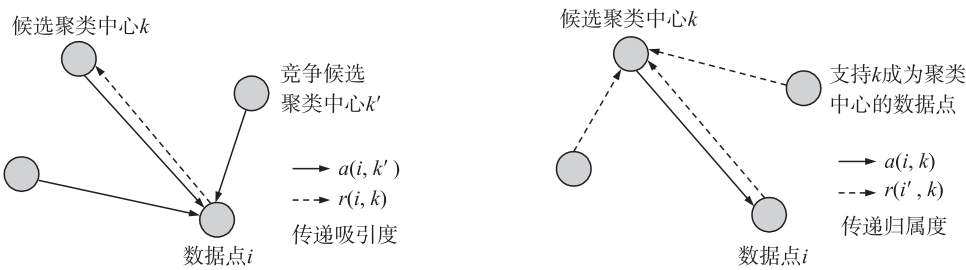


图 2 消息传递过程
Fig. 2 Process of sending messages

AP 算法输入的是 N 个数据点之间的相似度矩阵 S ,以矩阵 S 对角线上的数值 $s(k,k)$ 作为 k 点能否成为聚类中心的评判标准,称之为参考度 (Preference). $s(k,k)$ 的值越大, k 点竞争成为聚类中心的可能性就越大. 初始时,所有点的参考度设为相同的 P (通常取 S 的均值),归属感 $a(i,k)$ 和吸引度 $r(i,k)$ 均为 0,表明数据点 i 不属于任何一个类,数据点 k 也不是聚类中心. 之后, $r(i,k)$ 和 $a(i,k)$ 按照式(6)和(7)计算,当 $i=k$ 时, $r(k,k)$ 和 $a(k,k)$ 分别成为自吸引度和自归属感,其中自吸引度 $a(k,k)$ 的值按式(8)来计算. 自吸引度 $r(k,k)$ 和自归属感 $a(k,k)$ 越大,则表明 k 成为聚类中心的可能性也就越大.

$$r(i,k)=s(i,k)-\max_{k'\text{ s. t. }k'\neq k}\{a(i,k')+s(i,k')\},$$

(6)

$$a(i,k)=\min\{0,r(k,k)+\sum_{i'\text{ s. t. }i'\notin|i,k|}\max\{0,r(i',k)\}\},$$

(7)

$$a(k, k) = \sum_{i' \text{ s. t. } i' \neq k} \max \{0, r(i', k)\}. \quad (8)$$

AP 算法在每次迭代过程中不断地更新每一个数据点的吸引度和归属感,直到产生 k 个高质量的聚类中心,同时其余数据点也被分配到相应的聚类簇中.更新过程如式(9)和(10)所示,其中, $\lambda \in [0, 1]$ 为阻尼因子(一般取 0.5),当 λ 较大时,收敛速度较快,当 λ 较小时,收敛速度较慢.

$$r_{t+1}(i, k) = \lambda(s(i, k) - \max_{k' \neq k} \{a_t(i, k') + s(i, k')\}), i \neq k, \quad (9)$$

$$a_{t+1}(i, k) = \lambda a_t(i, k) + (1 - \lambda)(\min \{0, r_t(k, k) + \sum_{i' \notin \{i, k\}} \max \{0, r(i', k)\}\}), i \neq k. \quad (10)$$

迭代的目标就是寻求满足 $\arg \max_k (a(i, k) + r(i, k))$ 的数据点.

2.3 基于 AP 算法的图像聚类

基于 AP 算法的图像聚类步骤描述如下:

- (1) 提取图像 $i(i=1, 2, \dots, n)$ 的分块加权颜色直方图,获取特征向量;
- (2) 根据特征向量计算图像之间的相似度矩阵 S ,对每幅图像 i 的参考度 $s(i, i)$ 赋相同的初值 P ;
- (3) 初始化吸引度和归属感矩阵为 $\mathbf{0}$,设置适当的迭代次数 T 、阻尼因子 λ ;
- (4) 对于每幅图像 i ,迭代做以下计算 T 次:
 - (a) 计算吸引度 $r_t(i, k)$ 和归属感 $a_t(i, k)$, $k=1, 2, \dots, n$;
 - (b) 按照式(9)和(10)更新 $r_{t+1}(i, k)$ 和 $a_{t+1}(i, k)$, $k=1, 2, \dots, n$;
- (5) 对每幅图像 i ,计算 $e(i, i) = r(i, i) + a(i, i)$,如果 $e(i, i) > 0$,则 i 即为聚类中心.
- (6) 结束.

2.4 基于 k -means 算法的图像聚类

基于 K -means 算法的图像聚类描述如下:

- (1) 提取图像 $i(i=1, 2, \dots, n)$ 的分块加权颜色直方图,获取特征向量;
- (2) 初始化:从所有图像中任选 k 个作为初始聚类中心;
- (3) 重复执行以下过程,直到聚类中心不再改变:
 - (a) 计算图像 i 到各个中心的距离,根据最小距离原则将图像 i 划分到相应的簇;
 - (b) 更新每个簇的中心;
- (4) 结束.

3 实验与性能分析

3.1 性能评价标准

本实验采用 F -度量值来衡量聚类效果的好坏^[9]. F -度量值是信息检索中一种组合查准率和召回率指标的平衡指标.实验计算出来的聚类结果可检验每一幅图像在聚类后是否被划分为正确的类别以及同一个类别中是否包含了特定类别的图像.因此,可以计算每一个聚类 j 所属类别 i 的查准率 $P(i, j)$ 以及每一个聚类 j 所属类别 i 的查全率 $R(i, j)$.

设 n_i 是类别 i 的图像数目, n_j 是聚类 j 的图像数目, n_{ij} 是聚类 j 中隶属于类别 i 的图像数目,则查准率 $P(i, j)$ 和查全率 $R(i, j)$ 可分别定义为:

$$P(i, j) = \frac{n_{ij}}{n_j}, \quad R(i, j) = \frac{n_{ij}}{n_i}. \quad (11)$$

对应的 F -度量值 $F(i, j)$ 定义为:

$$F(i, j) = \frac{2P(i, j)R(i, j)}{P(i, j) + R(i, j)}. \quad (12)$$

全局聚类的 F -度量值定义为:

$$F = \sum_i \frac{n_i}{n} \max_j (F(i, j)), \quad (13)$$

其中, n 是图像集合中总的图像数目,通常, F -度量值越大,聚类效果越好.

3.2 实验结果分析

本文对 Corel 数据库(<http://wang.ist.psu.edu/iwang/test1.tar>) 中的图像数据进行了聚类分析实验,选择 Africa people、Seabeach、Buildings、Flowers 等 8 个类别,每类 80 幅共 640 幅彩色图像.实验按照本文所介绍的方法提取特征并计算相似度,最后分别用 AP 算法和 k -means 算法进行聚类,结果对比如表 1 所示.由表对比可知,AP 算法的聚类效果比 k -means 有明显提高.

4 结论

本文基于 HSV 颜色量化空间,利用分块加权颜色直方图方法提取图像的颜色特征,分别运用 AP 算法和 k -means 算法进行图像聚类分析实验.应用 AP 算法进行图像聚类不需要设定类别个数,聚类质量好,运行结果比较稳定,特别是在处理大数据量时效果尤为明显.而应用 k -means 算法进行图像聚类需要设定类别个数,且对于初始聚类中心的选择非常敏感,如果类别个数设定不恰当,则将对聚类结果有很大的影响.实验说明 AP 算法在处理图像这类大数据量数据时具有较高的 F -度量值,聚类效果更好.

[参考文献](References)

[1] Chen Yixin,Wang James Z,Robert Krovetz. Content-Based Image Retrieval by Clustering[C]//Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval. New York:ACM Press,2003:193-200.

[2] Sanjay Silakari,Mahesh Motwani,Manish Maheshwari. Color image clustering using block truncation algorithm[J]. International Journal of Computer Science Issues,2009,4(2):31-35.

[3] 黄祥林,沈兰荪. 基于内容的图像检索技术研究[J]. 电子学报,2002,30(7):1065-1071.
Huang Xianglin,Shen Lansun. Research on content-based image retrieval techniques[J]. Acta Electronica Sinica,2002,30(7):1065-1071. (in Chinese)

[4] 刘康苗,仇光,卜佳俊,等. 基于视觉和语义融合特征的阶段式图像聚类[J]. 浙江大学学报:工学版,2008,42(12):2 043-2 048.
Liu Kangmiao,Qiu Guang,Bu Jiajun,et al. Structured image clustering based on fusion of visual and semantic features[J]. Journal of Zhejiang University:Engineering Science Edition,2008,42(12):2 043-2 048. (in Chinese)

[5] Frey B J,Dueck D. Clustering by passing messages between data points[J]. Science,2007,315(5814):972-976.

[6] Han Jiawei,Micheline Kamber. Data Mining Concepts and Techniques[M]. 2nd ed. San Francisco:Morgan Kaufman Publishers,2006.

[7] 徐淑平,林福宗. 基于图像中心加权特征的图像检索[J]. 计算机应用与软件,2006,23(2):3-5.
Xu Shuping,Lin Fuzong. An image retrieval based on image center weighted feature[J]. Computer Applications and Software,2006,23(2):3-5. (in Chinese)

[8] 刘晓勇,付辉. 一种快速 AP 聚类算法[J]. 山东大学学报:工学版,2011,41(4):20-23.
Liu Xiaoyong,Fu Hui. A fast affinity propagation clustering algorithm[J]. Journal of Shandong University:Engineering Science Edition,2011,41(4):20-23. (in Chinese)

[9] 黄承慧,印鉴,侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度度量方法[J]. 计算机学报,2011,34(5):856-864.
Huang Chenghui,Yin Jian,Hou Fang. A text similarity measurement combining word semantic information with TF-IDF method[J]. Chinese Journal of Computers,2011,34(5):856-864. (in Chinese)

[责任编辑:严海琳]