

颜色特征的图像分类技术研究

李海峰¹, 杜军平²

(1. 北京工商大学 计算机学院, 北京 100037; 2. 北京邮电大学 计算机学院, 北京 100876)

摘要:研究了基于颜色的图像特征对于图像分类结果的影响. 给出了采用基于颜色位置分布特征进行分类的方法, 并与基于 RGB 直方图特征和基于 HSV 直方图特征的方法进行了比较. 分别采用随机森林、Boosting 算法和 MLP 神经网络 3 种分类方法进行图像分类, 建立了自然图像分类系统. 基于实验结果比较了随机森林、Boosting 算法和 MLP 神经网络 3 种分类方法的优缺点, 发现 Boosting 算法表现最好, 更加适合于图像分类.

关键词:图像分类; 颜色特征; Boosting 算法

中图分类号: TP391 **文献标识码:** A **文章编号:** 1673-4785(2008)02-0155-04

Image classification technology based on color features

LI Hai-feng¹, DU Jun-ping²

(1. School of Computer Science, Beijing Business and Technology University, Beijing 100037, China; 2. School of Computer and Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: This paper studies the effects of an image's color features on image classification. We developed a new classification method based on positional distribution of colors and compare it with other methods based on RGB and HSV histograms. The random forest, the Boosting algorithm, and the MLP neural network were applied respectively to classify images and a natural image classification system built up. The advantages and disadvantages of these three classification algorithms are discussed according to experimental results, showing that the boosting algorithm gives the best performance and is more suitable to image classification.

Keywords: image classification; color features; boosting algorithm

为了提高面向旅游业的图像搜索引擎的搜索质量, 实现对搜索结果进行筛选分类, 剔除与其他结果图像相似度较低的图像, 将会有助于提高搜索的结果质量. 由于搜索引擎的及时性要求, 需要一个比较简洁快速的方法进行图像分类. 图像分类技术是图像处理的基础, 图像特征提取是决定图像分类水平的关键. 图像特征提取即采用矢量数来描述图像的某种性质, 是图像处理的最基本问题之一. 图像底层视觉特征主要包括颜色、纹理和形状. 其中颜色特征比较容易提取, 处理速度也最快.

本文讨论基于颜色特征的图像分类技术, 并提出使用包含位置信息的颜色属性提高图像分类的质量的方法. 目前对多种类别图片进行分类的算法的

方法有很多种, 如神经网络、支持向量机、决策树等, 其分类效果各异^[1], 最近组合分类器和加强分类器也逐渐得到人们的重视. 本文将采用 3 种有代表性的算法, 即随机森林^[2]、Boosting 算法和 MLP 神经网络^[3-4]进行图像分类研究.

1 图像分类算法概述

本文主要讨论基于颜色的图像特征对于图像分类结果的影响. 为减少学习算法对结果的影响, 采用了 3 种学习算法进行比较和研究.

1.1 随机森林算法

随机森林算法通过自助法 (bootstrap) 生成多个树分类器, 它属于组合分类器^[5]. 其步骤如下: 从原始训练数据集 N 中重复随机抽取 k 个自助样本集, 每个自助样本集对应一棵分类树的训练数据集. 在一棵树的每个节点处, 从 M 个特征中随机挑选 m

收稿日期: 2007-08-07.

基金项目: 国家自然科学基金资助项目 (60773112); 北京市自然科学基金资助项目 (4082021).

通讯作者: 杜军平. E-mail: junpingd@bupt.edu.cn.

个特征 ($m \ll M$), 然后按照节点不纯度最小原则, 从这 m 个特征中选择一个特征进行生长. 节点 n 上的分类数据如果都来自于同一类别, 则此节点的不纯度 $I(n) = 0$. 不纯度度量方法采用 Gini 准则, 即假设 $P(j)$ 是节点 n 上属于 j 类样本个数占训练样本总数的频率, 则 Gini 准则表示为

$$I(n) = \sum_i P(i) P(j) = 1 - \sum_j P^2(j). \quad (1)$$

将生成的多棵分类树组成随机森林, 用随机森林分类器对新的数据进行判别与分类, 分类结果按树分类器的投票多少而定. 当生成一棵分类树时, 通常需要防止出现过拟合现象.

1.2 Boosting 算法

Boosting 算法是一种特殊的组合分类器方法^[6]. Boosting 方法使用基(弱)算法生成一系列的基分类器, 每个基分类器的训练依赖于上一次分类器的分类结果. 基分类器在训练集上的错误率用于调整训练样本的概率分布, 最终分类器通过单个基分类器的加权投票建立起来. Boosting 算法具有很多优点: 首先它简单易用, 除了迭代次数 T 以外, 不需要调节任何参数; 其次它不需要先验知识. 只要有足够多的数据以及弱学习器, 就能达到任意预测精度. Boosting 不易导致过学习, 所以不必担心训练的次 T . Boosting 算法也存在一些缺点: 过于依赖数据和基学习器, 对数据噪声很敏感. 如果基学习器太弱, 则不能达到任意高的精度. 在实验中采用了较为流行的 AdaBoost 算法^[7].

1.3 MLP 神经网络

MLP (multilayer perceptron) 是神经网络的一种实现. MLP 网络由感受层(S)、联想层(A)、响应层(R)构成. S、A、R 均由同类神经元构成. 反向传播(back propagation, BP)学习算法在 MLP 模型中得到广泛应用^[8]. 它把一组样本的输入、输出问题变为一个非线性优化问题. 网络开始训练时, 选用较小的随机互联权值与内部阈值, 通过反复加载训练样本并调整权值, 直到代价函数下降到可接受的容限值. 在试验中采用了 MLP 方法进行分类.

2 基于颜色的图像属性

颜色特征是图像特征中最为普遍、直观的特征, 它和图像中所包含的物体或场景相关, 对图像本身的尺寸、方向、视角的依赖性较小, 具有一定的鲁棒性^[9], 所以在图像检索中得到了广泛的应用. 也是本文研究的重点.

2.1 颜色直方图

颜色直方图是图像检索中最为常用的颜色特征, 它提取简便, 具有旋转不变性(rotation invariance)、尺度不变性(scale invariance)、平移不变性

(translation invariance), 而且对观测视角的变化不太敏感. 计算颜色直方图时需要将颜色空间划分成若干个小的颜色区间, 这个过程称为颜色量化. 颜色直方图的核心思想为: 采用一定的量化方法对颜色空间进行量化, 然后计算颜色落在每个小区间内的像素数量, 得到颜色直方图. 在本文的实验中, 采用图像的 RGB 空间的灰度直方图作为颜色特征进行分类, 只统计颜色的灰度属性, 并利用该属性进行分类处理.

2.2 改进的颜色直方图

面向硬设备的颜色模型与人的视觉感知有一定的距离. 例如给定一个颜色, 人们很难判断其中的 R、G、B 分量, 这时使用面向视觉感知的颜色模型比较方便. HSV 模型就是面向视觉感知的颜色空间(perceptual color spaces), 较为符合人类思维对颜色的理解、思考方式. 这一类模型还包括: HIS (hue, intensity, saturation) 等. 这些模型与人类颜色视觉感知比较接近, 而且独立于显示设备. 相对于其他颜色模型它更为准确地反映了人类视觉系统对色彩的理解方式^[10]. HSV 模型可以从 RGB 模型颜色空间进行转化. 假设 RGB 颜色 (r, g, b) , $r, g, b \in [0, 1, \dots, 255]$, 对应 HSV 空间 (h, s, v) 则

$$\begin{cases} v = \max(r, g, b), \\ r = \frac{v - r}{v - \min(r, g, b)}, \\ g = \frac{v - g}{v - \min(r, g, b)}. \end{cases} \quad (2)$$

令

$$h = \begin{cases} 5 + b, & r = \max(r, g, b) \text{ 且 } g = \min(r, g, b); \\ 1 - g, & r = \max(r, g, b) \text{ 且 } g = \min(r, g, b); \\ 1 + r, & g = \max(r, g, b) \text{ 且 } b = \min(r, g, b); \\ 3 - b, & g = \max(r, g, b) \text{ 且 } b = \min(r, g, b); \\ 3 + g, & b = \max(r, g, b) \text{ 且 } r = \min(r, g, b); \\ 5 - g, & \text{其他}. \end{cases}$$

得到 HSV 空间向量:

$$\begin{aligned} h &= 60 \times h, \\ s &= \frac{v - \min(r, g, b)}{v}, \\ v &= \frac{v}{255}. \end{aligned}$$

由于这样得到的属性众多, 而人眼无法分辨很小的分量之间的差别, 因此将 HSV 空间各个分量进行量化, 不会导致算法结果的差别太大, 而且可以减少计算量和提高计算速度. 本文的实验中, 将 HSV 3 个分量进行等间隔量化, 将色调 H 平均分为 x 份, 饱和度 S 平均分为 y 份, 亮度 V 平均分为 z 份, 得到 3 个标量 H, S, V , 其中 $H \in [0, x-1], S \in [0, y-1], V \in [0, z-1]$.

令属性矢量 $Z = H \times y \times z + S \times z + V$ 作为特征矢量,由此可得 $Z [0, x \times y \times z - 1]$,得到一维直方图,然后进行归一化,作为分类的颜色属性.对照 HSV 色表,发现当亮度 I ,饱和度 S 区分不是很明显,而对于色调 H 属性肉眼比较容易分辨出不同,因此试验中将 I 、 S 分量都均匀地分为 3 份,而对于 H 分量可均匀地分为 10 份.试验表明使用 HSV 空间较 RGB 空间有更好的分类能力.

2.3 考虑位置的颜色属性

前面 2 种直方图颜色属性都没有考虑颜色在画面中的位置,而颜色在图片中的位置对于人类识别一个图像起着非常重要的作用,例如画面中心的内容一般比画面边缘的内容更加重要.汽车图像和花朵的图像的背景可能都是大片的蓝天,这时直接使用全局的颜色属性就不容易区分出这是花朵图片还是汽车图像,而且容易出现过度拟合现象.如果这时考虑颜色在图像中的分布,就可以提高中心位置颜色属性的权重,从而达到更好的图像分类效果.本文采用一种考虑颜色位置的简单颜色属性,首先将图片分成 $x \times y$ 个等大的矩形小图片,即横向均匀分为 x 份,纵向均匀分为 y 份.为每个小图像进行编号,形成一个矩阵

$$\begin{bmatrix} a_{11} & \dots & a_{1x} \\ \dots & \ddots & \dots \\ a_{y1} & \dots & a_{yx} \end{bmatrix}.$$

对每个矩形小图像统计 RGB 各个分量以及灰度分量的和,然后根据小图像大小归一化得到 4 个标量,将这些标量依次排列起来作为该图像的颜色属性,共 $4 \times x \times y$ 个,然后依据这些属性对图像进行分类.在实验中取 $x = y = 3$,即将图像分为 9 张矩形图片.试验表明仅简单地将图像分为 9 块,就可以明显地提高基于颜色的分类质量.

3 实验结果及分析

3.1 图像库

采用百度图像搜索引擎搜索了关于建筑物、汽车和花朵 3 类图像,进行人工筛选,剔除不符合所在主题的图片,得到 544 张建筑物图像、913 张汽车图像、404 张花朵的图像作为实验的图像库.

3.2 实验环境

程序采用 Microsoft Visual Studio 2005 集成开发环境,使用 C++ 语言进行编写.测试机器是一台使用 AMD 2800+ CPU 和 1 GB 内存的 PC 机,操作系统使用 Windows XP SP2.

3.3 实验参数及结果分析

在测试 3 类基于颜色属性的实验中,采用如下方法处理图片:将所有图片中的 50 % 的图像作为训练样本集,使用另外 50 % 作为测试集.算法参数设

定如下:随机森林算法中,最大深度取 10,森林中树的最大数取 100,森林最小精度取 0.01.

在 Boosting 算法中,采用决策树算法作为弱分类器,弱分类器个数设置为 100.随着集合中树的数量的增加和信任度的增加,大部分的训练数据被正确分类,这些样本的权重不断的降低.具有较低相关权重的样本对弱分类器的训练有较低的影响,因此这些样本会在训练分类器时被排除在外,而不对分类器造成较大影响.在算法实现中,样本系数小于 0.05 时被排除.在使用灰度颜色直方图实验中,使用灰度的 256 个分量作为属性.对这些分量按照图像大小进行归一化,每幅图像共由 256 个属性描述.实验结果如表 1 所示.

表 1 基于灰度颜色直方图颜色属性的学习结果
Table 1 Color attribute results based on gray histogram

算法	图片集 合数量 / 例	样本集 合符合 数量 / 例	样本集 合正确 率 / %	测试集 合正确 数量 / 例	测试集 合正确 率 / %
Random Forest	1 861	507	54.5	472	50.7
Boosting	1 861	930	100.0	559	60.1
MLP	1 861	927	99.7	429	46.1

可以看出测试集合正确率不高,由于使用的属性比较多,处理速度也比较慢.这说明基于 RGB 颜色统计的属性作为图像分类属性区分能力不高.在 HSV 颜色直方图实验中,参数设定为 $x = 10, y = 3, z = 3$,即每张图像由 90 个属性描述,实验结果如表 2 所示.

表 2 基于 HSV 颜色直方图颜色属性的学习结果
Table 2 Color attribute learning results
based on HSV color histogram

算法	图片集 合数量 / 例	样本集 合符合 数量 / 例	样本集 合正确 率 / %	测试集 合正确 数量 / 例	测试集 合正确 率 / %
Random Forest	1 861	881	94.7	697	74.9
Boosting	1 861	930	100.0	693	74.5
MLP	1 861	930	100.0	606	65.1

由结果可知使用 HSV 颜色直方图的方式明显优于 RGB 颜色直方图的方式,MLP 算法出现了一定程度上的过度拟合现象.考虑位置时实验中参数 $x = 3$,即每幅图像分为 9 个小图像,最后每幅图像

由 36 个属性描述,结果如表 3 所示.

表 3 基于颜色位置分布的颜色属性学习结果
Table 3 Color attribute learning results based
on color position distribution

算法	图片集 合数量 / 例	样本集 合符合 数量 / 例	样本集 合正确 率 / %	测试集 合正确 数量 / 例	测试集 合正确 率 / %
Random Forest	1 861	874	94. 0	722	77. 6
Boosting	1 861	877	94. 3	729	78. 3
MLP	1 861	918	98. 7	651	69. 9

由结果可以看出基于颜色位置分布的颜色属性,可以得到较好的学习效果. 由于属性比较少,而且计算简单,所以试验的执行速度也比较快.

4 结束语

单纯依赖颜色的一些统计属性进行图像分类,可以有一定的分类能力,但是也有一定的局限性,这是由于图像的最基础视觉属性不仅仅是颜色,还包括形状和纹理. 在使用颜色直方图时,HSV 模型较 RGB 模型有更好的识别能力,这是因为 HSV 模型更接近人的视觉感觉,而 RGB 模型是一种设备描述模型,因此从人类识别图像的角度进行图像分类,可以得到更好的效果. 仅仅粗略考虑颜色在图像中位置的分布,同时采用最简单的颜色统计,亦可以大幅度提高图像的分类能力. 由此可以推断,如果采用更精细的图像颜色属性的位置信息,可以更好地提高图像分类质量. 对于 3 种学习方法,可以发现 Boosting 方法表现最好,更加适合于图像分类. 颜色属性是图像属性的最基本属性,提取最为方便快捷,使用此属性可以达到一定的分类能力,因此可应用于要求不高的分类场合,例如图像搜索结果的筛选.

参考文献:

[1]章毓晋. 图像分割[M]. 北京:科学出版社, 2001.
[2]贾富仓,李 华. 基于随机森林的多谱磁共振图像分割[J]. 计算机工程, 2005, 31(10):159-161.
JIA Fucang, LI Hua. Multi-spectral magnetic resonance Image segmentation using random forests[J]. Computer Engineering, 2005, 31(10):159-161.

[3]吕 坤,高 珊. 一种基于改进的 BP 神经网络算法的布匹瑕疵分类器[J]. 微电子学与计算机, 2006, 23(3):88-90.
LU Kun, GAO Shan. A fabric flaws classifier based on improved BP neural network algorithm[J]. Microelectronics & Computer, 2006, 23(3):88-90.
[4]BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1):5-32.
[5]SCHAPIRE R E. The strength of weak learnability[J]. Machine Learning, 1990, 5(2):197-227.
[6]SCHAPIRE R E. A brief introduction to boosting[C]// Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Stockholm, 1999:1401-1406.
[7]FABRIZIO S, ALESSAMDRO S, NICOLA V. An improved boosting algorithm and its application[C]// Proceedings of the 2000ACM CIKM International Conference on Information and Knowledge Management. Mclean, VA, USA, 2000:78-85.
[8]HAN Jiawei, KAMBER M. Data mining: concepts and techniques [M]. 北京:机械工业出版社, 2001.
[9]杨再华,李玉和,李庆祥. 基于边缘特征提取的图像清晰度评价函数[J]. 计算机工程与应用, 2005 (10):35-36.
YANG Zaihua, LI Yuhe, LI Qingxiang. Image definition criterion based on edge gradient feature[J]. Computer Engineering and Applications, 2005 (10):35-36.
[10]MANJUNATH B S, OHM J R, VASUDENVAN V V, et al. Color and texture descriptors[J]. IEEE Trans on CSVT, 2001, 11(6):703-715.

作者简介:



李海峰,男,1980 年生,硕士研究生,主要研究方向为模式识别、智能信息系统等.



杜军平,女,1963 年生,教授,博士生导师,中国人工智能学会常务副秘书长,中国自动化学会智能自动化专业委员会副秘书长,中国旅游信息标准技术委员会主任委员,主要研究方向为数据挖掘、Agent 理论与技术、智能信息处理、旅游智能信息系统等. 近年来完成科研项目 20 余项,发表学术论文 90 余篇,出版著作 3 部.