# 高级数据结构课程实验一报告

洪方舟 2016013259 hongfz16@163.com
李帅 2016013270 lishuai16THU@163.com
周展平 2016013253 zhouzp16@163.com

## 摘要

本实验对 RTree 性能进行测试，并且利用 RTree 对 5000 张图片提取的颜色直方图特征、颜色矩特征进行检索并使用多种评价方式分析以上特征的检索效果，最后对 RTree 进行改进以提升性能，优化检索结果。

## 关键词

图像检索；RTree；颜色直方图特征；颜色矩特征；

## 1. 实验系统介绍

本次实验系统主要包括RTree、特征提取器、测试部分。由 RTree 向测试函数提供标准接口，特征提取器向测试部分提供图片特征。下面对三个部分分别进行详细介绍。

### 1.1 RTree 标准化接口

本次实验中使用RTree为助教提供的c++版本的模板，做过必要的修改后可以完成本实验测试任务。为方便测试，本实验对测试中使用到的 RTree 接口做标准化规定，包括插入操作，查询操作，获取节点访问次数。

### 1.2 特征介绍

#### *1.2.1 颜色直方图特征*

颜色直方图是图像的一种颜色特征。图像的每种颜色都不是单一的，可以由几种单一独立的颜色混合而成。在给定的色彩空间上，将每个像素的颜色沿着不同的通道（每种通道表示一种单一的颜色分量）分离开，统计每种颜色分量的总数或者占所有颜色的比例，用直方图将统计结果表示出来即为颜色直方图。颜色直方图是图像的一种全局特征，他对于图像旋转、缩放、模糊等物理变换并不敏感，因此可以用来衡量和比较不同图像的全局差。但是这种全局特征也会导致像素点间的位置特征丢失，例如几幅整体色调相近的图片，但是颜色的分布完全不一样，这种情况下颜色直方图无法对其进行区分。

本实验中使用在 RGB 颜色空间下提取颜色直方图作为特征。RGB 颜色空间将颜色分解为红色（Red）、绿色（Green）、蓝色（Blue）三个分量，其中每个分量安亮度可分为 256 个等级。通过对每个颜色分量的每个亮度值的像素点数进行统计，便可得到3个256维的特征。之后再对每个通道的亮度值进行划分，将一定范围内的像素值归为一类，便能将 256 维特征压缩到更小的维数。本实验中使用的特征维数是3*3、3*5和3*8。3*3的特征将每个通道的亮度值划分为0-84,85-169,170-255三个范围，其他维数的特征划分与此类似。

#### *1.2.2 颜色矩特征*

矩特征是一类重要的图像特征，颜色矩特征(color moment)是一类常用的矩特征，优势在于维度较低，且能够较为充分表达图像的颜色特征。具体来说，颜色矩特征是一个包含 9 个维度的向量：

$$\text{color moment} = (\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, s_1, s_2, s_3)$$

对于一幅$m * n$的图像，3 个颜色通道分别占用 3 个特征，分别为：

$$\mu = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$$

代表颜色通道的均值(mean),

$$\sigma = \left(\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\mu)^2\right)^{\frac{1}{2}}$$

代表颜色通道的方差(variance),

$$s = \left(\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\mu)^3\right)^{\frac{1}{3}}$$

代表颜色通道的斜度(skewness)。

### 1.3 测试方法

#### *1.3.1 Problem1*

经过前期预实验发现，测试结果在较少采样量下无法呈现一致的规律，因此实验一中对于每一种维数的特征，每一个查询集大小，重复测试 20000 次，取平均结果。为了观察到较为准确的结果，本实验中对维数的采样范围为(4,24)内所有整数，对查询集的采样范围为(1000,5000)，每隔500取一点。实验中对于每个维数，每个查询集大小测试两项数据，一项是在固定的查询范围下（每一维度上查询范围均一致）的磁盘读取次数（本实验中使用节点访问次数来近似获取该数据），另一项是平均每查询出 100 个结果所需要的磁盘读取次数。由于本实验所使用的 RTree 的实现中没有返回节点访问次数的接口，因此需要对原实现做相应修改。

#### *1.3.2 Problem2*

对于每种不同的特征，记录多组数据，包括每次查询返回的结果数量、准确率(Precision)、召回率(Recall)，通过下式计算 F 值(F-Measure)：

$$F = \frac{(\alpha^2+1)Precision \times Recall}{\alpha^2(Precision + Recall)}$$

在本实验中取$\alpha = 1$，此时 F 值为准确率与召回率的加权调和平均值，能够较好的评价查询效果。以查询返回结果数量为 x 轴，F 值为 y 轴，做出曲线进行分析。

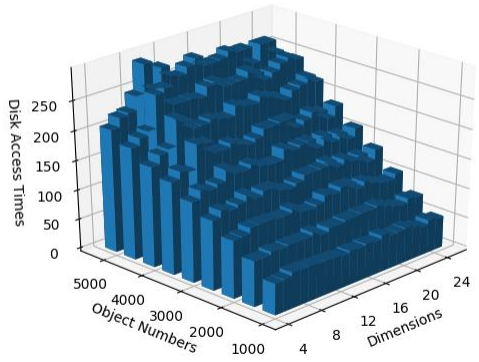
### *1.3.3 Problem 3*

对于每次查询出来的结果集合，对每个结果的特征向量与查询图片的特征向量的欧式距离进行排序，取距离最近的一百张图片，采取如下的打分规则，对于第 i 张图片，如果该图片与查询图片属于同一类，则在总得分加上 100-i 分，否则减去 100-i 分。多次测试取平均得分作为当前测试特征的得分。

## 2. Problem 1 测试结果及分析

结果如立体直方图所示，上方图片 z 轴坐标为固定每一维度查询范围的磁盘读取次数，下方图片 z 轴坐标为每次查询中平均取回 100 张图片所需磁盘读取次数。下面分别对两个结果进行分析。

### 2.1 固定查询范围

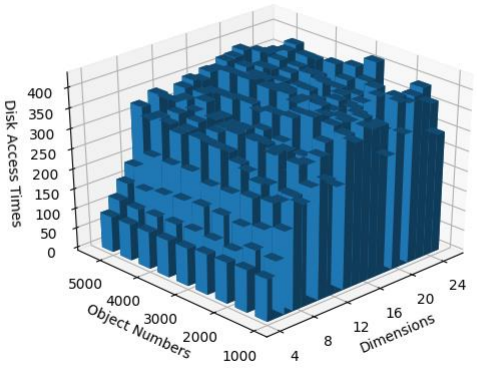



通过立体直方图可以直观的看出，在相同的特征维数下，随着查询集增大，平均磁盘访问次数增多。分析原因如下：相同维数下，固定查询范围，则对于某个特定的查询点，周围点的密度会随着查询集的增大而增大（该性质由测试过程中随机选取图片插入 RTree 的过程保证，由于随机选择，点的分布相对较为均匀，因此随着插入的点的数量增大，点的密度将会随之增大），因而查询返回的结果数量将会增大，从而需要访问磁盘的数量会增多。

对于相同查询集大小的情况下，随着维数增加，磁盘访问次数并没有呈现出一致的总体变化规律，但是在较低维的范围内（4-8 维），随着维数增加，磁盘访问次数增加。分析原因如下：较低维，例如 4 维特征并不能很好的表示相应的图片，导致相应的特征空间中相同类型的图片并没有很好的聚集在一起，而是整体分布较为松散，因此在固定查询范围的情况下，随着维数增大，也即特征空间中类别聚集程度增大，每次取出的结果数量增大，自然所需要的磁盘访问次数也会增大；而当维数增大到一定的程度，例如 8 维及更高的情况下，再增加维数对于特征空间中聚集程度的增加已经没有非常明显的帮助了，因此在较高维数的范围内没能观察到随维数增加磁盘查询次数增多的现象。



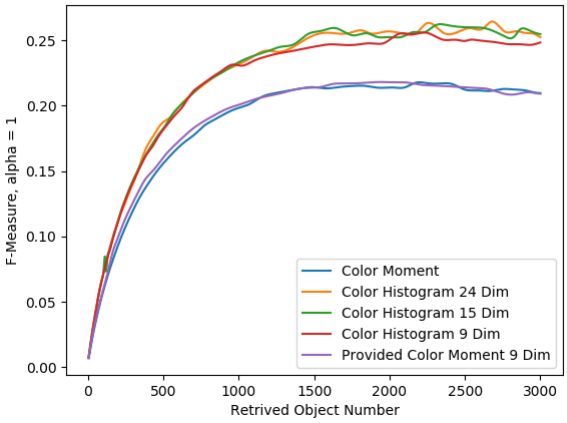
### 2.2 平均取回 100 张图片




通过立体直方图可以直观地看出，在相同的查询集大小下，随着特征维数增高，平均每取回 100 个查询结果所需要的磁盘访问次数增多。分析原因如下：随着特征维数的增高，RTree 中矩形框重合的数量增多（该性质可作如下简要证明，假设特征为 n 维，两个在每一维上取值均满足平均分布的 n 维矩形框不重合的概率 $P = p^n$，其中 p 为在一直线段上随机取两直线段重合的概率，且有 $p < 1$，因此随着特征维数的增高，n 维矩形框重合的概率增大），在对于每个节点的查询中平均需要访问更多的子节点，因此磁盘访问次数增多。

在维数相同的情况下，从测试结果来看，无法观察到平均取回 100 个结果所需要访问磁盘的次数与查询集大小的关系，分析认为上述两个量无明显关联。原因如下：随着查询集增大，特征空间中查询图片周围点的密度增大，因而固定查询范围的情况下，磁盘访问次数会增多，但是在平均取回 100 个结果的情况下，将查询范围的影响排除，其他因素如 RTree 节点中矩形重合情况，在相同维数下基本维持一致，因此认为平均取回 100 个结果所需要访问磁盘的次数与查询集大小无明显关联。

## 3. Problem 2 测试结果及分析



从查询结果所作的曲线图可以看出，总体上颜色直方图特征的查询效果好于颜色矩特征。对于颜色直方图特征，随着查询点集的不断增大，维数为 9 维、15 维、24 维的颜色直方图作为特征的查询，F 值均不断提高，最终三种维数的 F 值均稳定在 0.25 附近。这个准确率并不算太高，对此的分析正如上文所述，颜色直方图是全局的颜色统计特征，不能很好的表达颜色的空间分布，因此对于图形的形状不能很好的分辨，导致了查询准确率较低。

另一方面，随着查询点集的增大，三个维数的 F 值始终非常相近，没有拉开差距。对于这个问题的分析是，三个特征看似维度不一，但实际上都是由同样的 3*256 维向量经过不同的划分而得，因此三个特征包含的信息十分相近，故即使维数不同三者的查询结果也都较为相近。

# 4. Problem 3 测试结果及分析

在本次实验中，我们比较了 5 种特征的检索结果，分别是：9 维、15 维、24 维的 Color Histogram 特征，提供的 Color Moment 特征以及我们自己实现的 Color Moment 特征。我们采用自己设计的评分策略对检索效果的好坏进行衡量，结果在下面的表格中。

**Table 1. 特征对检索评分的影响**

| Graphics | Top |
|---|---|
| Color Histogram (dim=9) | -295.098 |
| Color Histogram (dim=15) | -258.124 |
| Color Histogram (dim=24) | -275.913 |
| Color Moment (provided) | -2081.56 |
| Color Moment (extracted) | -2434.01 |

分析表中数据，可以得到如下结论：

（1）Color Histogram 特征在检索效果上明显优于 Color Moment 特征。

（2）维度与 Color Histogram 特征的检索效果之间没有显著的相关性。

（3）各类特征的得分均小于 0，说明检索效果均较差。

# 5. Optional 部分实验

## 5.1 题目(b)实验

R*-Tree 是 RTree 的一种变体，它的主要改进在于：

（1）在插入过程中选择子树以及节点分裂的时候，不仅仅考虑传统 RTree 所考虑的矩形区域面积的增量，还同时考虑了矩形区域的重叠问题。此改进在理论上可以提高查询效率，因为重叠的减少意味着被重复查询的节点可能会减少 。

（2）在插入过程中使用强制重新插入的技术，在理论上可以尽量避免 RTree 在删除后不平衡的问题，通过提升平衡度来提高查询效率。

## 5.2 题目(c)实验

原 RTree 中进行近邻查询时使用的是 L2 距离（欧式距离）进行排序，因此考虑使用 L1 距离（曼哈顿距离）以及 L3 距离作为近邻查询中度量相关度的函数，实验结果如下，表格中数据为采用 Problem3 的评分算法得出的分数。

| 特征名称 | L1 距离 | L2 距离 | L3 距离 |
|---|---|---|---|
| Color Histogram (dim=9) | -925.549 | -963.701 | -1076.72 |
| Color Histogram (dim=15) | -690.243 | -706.103 | -815.982 |
| Color Histogram (dim=24) | -361.384 | -380.994 | -474.873 |
| Color Moment (provided) | -2105.59 | -2124.92 | -2253.74 |
| Color Moment (extracted) | -2456.85 | -2427.04 | -2641.41 |

从表格中可见，采用 L1 距离或者 L3 距离对于颜色矩特征的查询效果无明显提升，甚至有倒退的情况。但是 L1 距离对颜色直方图特征的查询效果提升明显，而 L3 距离则使之查询效果变差。分析可能的原因如下：颜色直方图特征中每一维数据实际上是某一颜色范围像素点数统计量，因此对该特征进行高次方的运算并没有实际意义，因而 L1 距离具有最好的效果，而阶数越高则效果越差。

# 6. REFERENCES


[1] Guttman, A. (1984). R-trees: a dynamic index structure for spatial searching. ACM SIGMOD International Conference on Management of Data (Vol.14, pp.47-57). ACM.

[2] Beckmann, N., Kriegel, H. P., Schneider, R., & Seeger, B. (1990). The R*-tree: an efficient and robust access method for points and rectangles. ACM SIGMOD International Conference on Management of Data (Vol.19, pp.322-331). ACM.