DOI:10.13718/j.cnki.xsxb.2019.07.009

基于 PCA-二叉树的大规模图像索引技术研究①

周雪梅, 潘 多

四川工商职业技术学院 信息工程系, 四川 都江堰 611830

摘要: 针对大数据数据库中图像索引中维度灾难问题, 该文提出一种基于云的大规模图像检索技术, 该方法创新性地将主成分分析法和二叉树引入到图像检索技术中, 首先采用尺度不变特征变换和加速鲁棒特征描述符作为帧特征, 面对大规模维度特征, 将主成分分析法对帧特征进行降维, 并使用二叉树表示降维后的特征, 以加速研究阶段并减少存储空间, 最终实现图像检索. 实验表明: 该文方法在降维 70%的条件下, 搜索精确率/召回率(Precision/Recall, *PR*)值能够达到传统方法 20%降维条件下的 *PR* 值, 并且在搜索时间上, 该文方法与正常搜索相比, 搜索速度得到 30%~50%的提升.

关 键 词: 大数据; 大规模图像索引; 主成分分析; 二叉树; 尺度不变特征变换

中图分类号: TP311 **文献标志码**: A **文章编号**: 1000-5471(2019)07-0057-06

随着云计算的发展, 用于不同领域的多媒体数据库中产生了越来越大量的图像、视频和声音等数据[1-2], 为了快速访问这些数据, 必须对所有这些数据进行索引, 索引图像[3-4]代表与计算机视觉有关的各种领域的必要工具, 如视频监控和运动分析, 索引过程成为与大数据领域相关的热点问题[5]. 对于大型数据库内图像, 通常提取高维特征来精确描述图像内容, 如果直接处理这些高维数据, 可能会导致维度灾难问题, 降低了索引算法的性能[6].

大规模图像检索是有效利用大数据的关键技术领域之一, 基于内容的图像检索[7]已经成为流行的方法, 该类方法通过图像处理和计算机视觉算法自动检测和提取图像的视觉特征(全局和局部特征), 然后与存储在数据库中的一组图像特征进行比较. 最后向用户显示和查询具有相似特征的图像列表. 在我们的案例中, 结果是与查询具有相似功能的视频列表, 降维是用于克服这些问题的有效方法之一. 文献[8]中提出一种基于矩阵指数嵌入来推断高维数据低维表示的降维框架, 在该框架中矩阵指数可以通过特征相似度矩阵上的随机游走来粗略解释, 并且因此更加鲁棒. 文献[9]中提出两种 k 均值聚类的降维方法, 一种是基于随机投影的 k 均值聚类降维方法, 另外一种是基于奇异值的降维方法, 两种方法都能够准确特征提取, 并且降低了时间复杂度. 另外还有其他方式用于降维, 如线性判别分析(Linear Discriminant Analysis, LDA), 支持向量机(Support Vector Machine, SVM)等方法.

目前对于图像检索的研究正在进行, 文献[10]中提出了一种针对大数据设计的快速图像检索方法, 该方法首先针对每个图像获得特征向量, 之后编码图像特征向量并将它们放入数据库, 这可以优化特征结构, 最后, 使用相应的相似性匹配来确定检索结果. 文献[11]中提出了针对大规模图像数据库中基于尺度不变特征变换(Scale Invariant Feature Transform, SIFT)特征和基于内容的图像检索方法, 该系统从每个原始图像中提取 SIFT 特征向量, 根据视觉相似性计算结果将视觉相似的图像返回给用户, 其创新之处在于引入 SIFT 描述符来表示图像的可视内容, 然后利用距离比作为阈值来控制匹配特征点的个数. 文献[12]提出了一种新的基于内容的图像检索方法, 改进可扩展词汇树图像检索方法, 解决了传统基于内容的

① 收稿日期: 2018-05-24
基金项目: 教育部科技发展中心产学研创新基金课题 (2018A03007).
作者简介: 周雪梅(1969-), 女, 硕士, 副教授, 主要从事虚拟现实技术, 计算机应用及大数据可视化研究.

图像检索技术通过特征向量来表达每幅图像在大规模图像检索中的精度和时间问题.

本文在研究了已有图像检索方法的基础上，针对大规模图像数据库检索中的特征维度灾难，以及搜索时间长的问题，提出了一种高效率的大数据数据库图像索引方法. 该方法首先将 SIFT 和加速鲁棒特征(Speeded up Robust Features，SURF)提取为图像特征，然后采用 PCA 降维方法来减小这些特征的尺寸，最后，提出了一种基于二叉树的图像存储结构表示方法，以加快索引时间.

1 本文方法

本文方法的主要目标是在系统中对图像进行索引，允许用户对一组图像进行研究. 本文方法分为 3 个阶段，第 1 阶段是特征提取；第 2 阶段是应用 PCA 作为降维方法来减少特征维数；最后是二值树的生成，中值是 SIFT 和 SURF 特征的每个节点. 本文方法总体结构如图 1 所示.

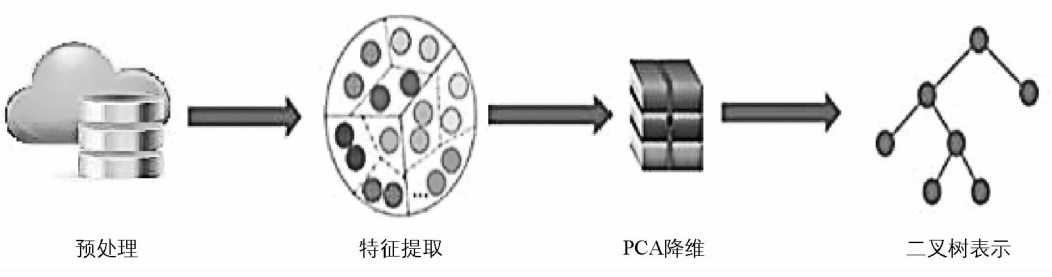

图 1 本文方法总体结构图

1.1 预处理阶段

首先，将高斯模糊应用于所有图像，以减少 SIFT 或 SURF 应用于这些图像时的关键点数量，在实验中应用了几个过滤器，实验表明 SIFT 和 SURF 的最佳滤镜是高斯模糊，高斯模糊是一种图像模糊过滤器，使用高斯函数来计算应用于图像中每个像素的变换，一维高斯函数的方程为

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} \tag{1}$$

其中，x 是在水平轴上与原点的距离，σ 为高斯分布标准差，公式(1)用正态分布计算每个像素的变换. 对于二维空间，每个维度都有一个

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{2}$$

其中，x 是在水平轴上与原点的距离，y 是垂直轴上与原点的距离，σ 为高斯分布标准差. 算法 1 中显示了这个预处理过程.

Pretreatment: *jpeg file*: *Set of images*;
  $\varnothing \rightarrow jpeg\_file$;
  ReadNb - *images_i*;
  for *j*: 1 *to Nb -images* **do**
    Get Current Image(j) → image_j;
    GaussianBlur(*image_j*) → $image\_GB_j$;
    $jpeg\_file \cup image\_GBj \rightarrow jpeg\_file$;
  end for
return jpeg_file;

算法 1 接受一组图像作为输入，计算每个图像的高斯模糊并存储这些图像，这个输出将被用作下一步的输入.

1.2 特征提取阶段

在特征提取阶段，本文使用 SIFT 和 SURF 描述符，描述符允许提取兴趣点图像. 尽管 SIFT 和 SURF 描述符的工作方式不同，但这 2 种情况下的输出都提供了兴趣点周围的邻域表示作为描述符向量，描述符向量可以比较或匹配从其他图像提取的描述符. 在本文中使用 OpenCV 作为库，并使用一些函数来计算

SIFT 描述符，本文采用的 SIFT 描述符是 $n\times128$ 矩阵. SURF 描述符基于 SIFT 描述符，本文应用的 SURF 描述符是 $n\times64$ 矩阵.

为了在 SIFT 和 SURF 特征之间进行比较，需要在几个度量标准中定义一个相似度量度：FLANN 匹配器和 Brute-Force 匹配器.

FLANN 是快速最近邻搜索包(Fast_Library_for_Approximate_Nearest_Neighbors，FLANN)的简称，它是一个对大数据集和高维特征进行最近邻搜索算法的集合，而且这些算法都已经被优化过了. 对于大型数据集，此度量比 Brute-Force Matcher 快，经验证 FLANN 比其他的最近邻搜索软件快 10 倍. FLANN 度量使用分层 k 均值树进行通用特征匹配，这使得 SURF 具有固有的优势，因为二元特征不易扩展到分层 K 均值. 图 2 显示了带有 SURF 描述符的基于 FLANN 的匹配器.

图 2　使用 SURF 描述符基于 FLANN 匹配器的匹配

Brute-Force 匹配器是使用 SIFT 特征描述符进行图像比较更简单的方法. 它首先采用一个特征描述符，并通过使用一些距离计算将它与第 2 组中的所有其他特征进行匹配，并返回最接近的特征. 图 3 显示了与 SIFT 描述符的强力匹配.

图 3　使用 SURF 描述符基于 Brute-Force 匹配器的匹配

1.3　降维阶段

无监督和有监督方法都可以应用于减少图像特征尺寸，在本文中使用降维方法 PCA. PCA 是基于各种方程的数学程序. 首先，按照等式计算协方差矩阵(3) 和(4).

$$u=\frac{1}{N}\sum_{i=1}^{n}x_i \tag{3}$$

$$Cov=\sum_{i=1}^{n}x_i(x_i-u)\,(x_i-u)^T \tag{4}$$

其中，$x_1, x_2, \cdots, x_n$ 是一组 N 维的特征值，u 是所有图像特征的均值，然后可以按照式(5)计算 W_{PCA}.

$$W_{\mathrm{PCA}} = \arg\max \mid V^T CovV \mid \tag{5}$$

本文已经提出应用PCA降维10%～90%的范围. 使用 $N_{compression} = 128 - (N \times 128/100)$ 来计算SIFT的压缩维度，$N_{compression} = 64 - (N \times 64/100)$ 来计算SURF的压缩维度，其中 N 在10～90之间.

1.4 图像存储结构阶段

在这一阶段中，本文选择了二叉树作为上一步生成压缩图像的存储结构. 这种结构能够加速更多的研究阶段，因为它与其他结构相比是最简单的结构，即使在实现层面，二叉树也可以用简单数组表示，并且数组的操作非常简单快捷. 图4给出了二叉树的一个例子.

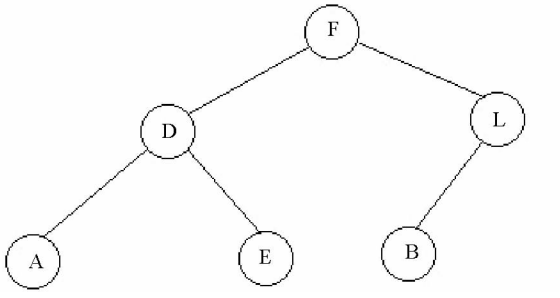

图4 二叉树结构

左侧的所有节点都小于根节点，右侧的所有节点都大于根节点. 在二叉树中，通常节点具有无法在它们之间进行比较的值，但在本文方法中用矩阵表示SIFT或SURF特征，并且可以在2个矩阵之间进行比较，所以本文提出了一种算法，能够在2个矩阵之间进行比较. 为了比较两个SIFT或SURF特征，本文使用相似性度量，结果是[0 1]之间的正常值，定义了两个度量区间[0 0.5]和[0.6 1]，分别表示低于根节点的每个节点区间和高于根节点的每个节点区间. 在算法2中给出详细步骤.

Algorithm 2 *Binary_Tree_Generation*(*S_Features*)
Require: *S_Features*: Set of features;
Ensure: *Table_Root*: Table contains the position of the nodes.
root ←random(features);
root →*Table_Root*;
∅→*Features_Left*;
∅→*Features_Right*;
length(*S_Features*)→*Nb-features*;
if *Nb* − *features* = 1 **then**
for *i* : 1 to *Nb* − *features* **do**
if (*feature_i*≠*root*) **then**
Distance(*feature_i*, *root*)→*Dist* ;
if (*Dist*≤0.5) **then**
Features_Left ∪ *feature_i* →
Features_Left;
else
Features_ Right ∪ *feature_i* →
Features_ Right;
end if
end if
end for
Binary_Tree_Generation(*Features_Left*);
Binary_Tree_Generation(*Features_Right*);
else
Break;
end if
return *Table_Root*;

在算法 2 中，选择一个具有随机函数的根，通过使用相似性度量之一来计算根和特征 i 之间的距离，然后决定是否将该功能置于左侧或右侧. 应用算法 2 直到获得完整的二叉树，二叉树的一个限制是所有节点都在左侧或右侧，在这种情况下，有必要平衡如图 5 所示的二叉树.

已经通过实验方法应用了平衡系统，并获得最佳结果.

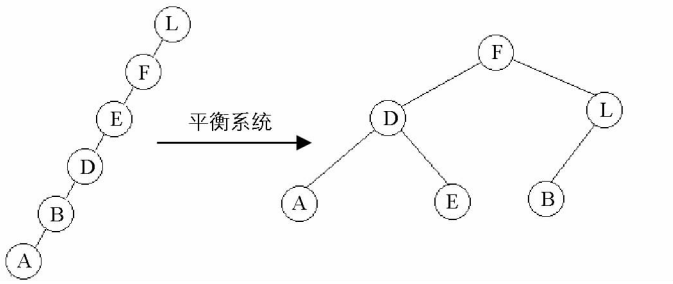

图 5　二叉树平衡系统

2　实验结果与实验

在机器学习中，ROC(Receiver Operator Characteristic)曲线被广泛应用于二分类问题中来评估分类器的可信度，但是当处理一些高度不均衡的数据集时，PR(Precision-Recall)曲线能表现出更多的信息. 为了评估本文方法，本次实验使用精确率/召回率作为评价指标，精确率/召回率公式见式(6)和式(7).

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

其中，TP 表示把正类预测为正类，FN 表示把正类预测为负类，FP 表示把负类预测为正类.

为了对比各种降维方法对大数据数据库图像检索性能的影响，本文使用名为 MIR Flickr 图像数据集进行实验，该数据集包含 100 万张图像，运行环境为云(Ubuntu)和 UMONS 集群上的操作系统.

图 6 显示了使用不同压缩比(20%和 70%)时精确率/召回率度量的演变.

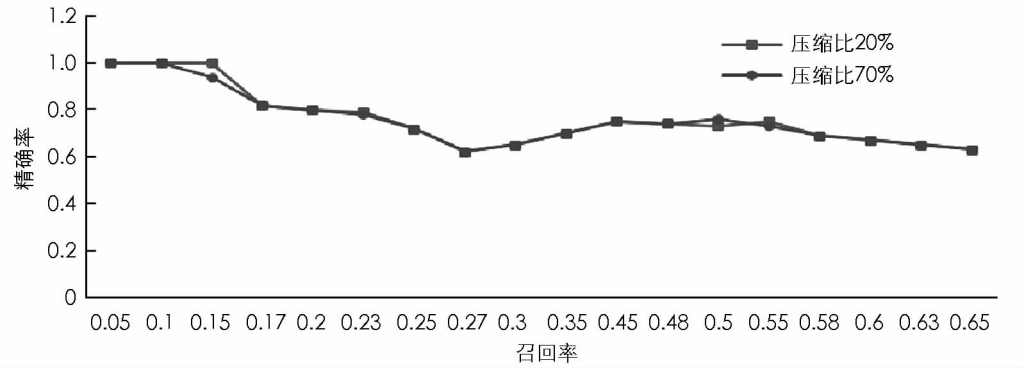

图 6　精确率/召回率曲线

从图 6 中可以得到，使用 70%的压缩比可以与使用 20%的压缩比获得几乎相同的 PR 曲线. 因此，本文方法的压缩不会对精度产生负面影响，这种压缩对于减少计算时间非常有效.

图 7 比较了本文方法与不使用二叉树搜索方法的搜索时间，基于二叉树的方法获得索引时间更少，时间减少 30%～50%. PCA＋SURF 特征存储需要较低的空间，且搜索用时最少，仅为 69 s，因为 SURF 索引的比例在 70%以内，这个比例(70%)所需的存储空间是使用 20%压缩时所需空间的一半.

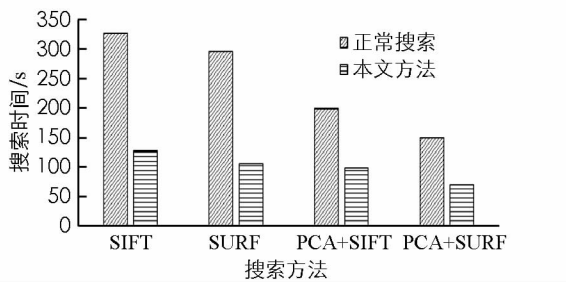

图 7　传统方法与本文方法比较

3　结　　论

针对大数据数据库图像索引维度和时间问题，本文提出了基于 PCA-二叉树的大规模图像索引方法，通过应用 PCA 和二叉树表示数据来显示和评估通过不同策略在大规模图像中降维获得的实验结果. 实验结果表明，与 20%的压缩比相比，70%的压缩比得到的 PR 曲线几乎相同，因为本文方法在索引时间和存储空间 2 个方面都有所提升. 此外，与传统图像索引方法

比较，本文方法能够减少索引时间，最高减少50%. 在接下来的工作中，计划通过使用Hadoop和HDFS作为分布式文件系统和MapReduce作为编程模型来改进本文方法，以便在数据增加时应用并行化处理.

参考文献：

[1] WU X D, ZHU X Q, WU G Q, et al. Data Mining with Big Data [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(1): 97-107.

[2] 陈小玉，李晓静，马海英. 一种面向大数据的快速自动聚类算法 [J]. 计算机应用研究，2017，34(9)：2651-2654,2658.

[3] 余 琨，伍孝金. 区域相关融合纹理特征FDPC图书馆文档图像检索 [J]. 西南师范大学学报(自然科学版)，2017，42(7)：91-98.

[4] CHAKER A, KAANICHE M, BENAZZA-BENYAHIA A, et al. An Efficient Image Retrieval Method under Dithered-Based Quantization Scheme [C]. Budapest: 24th European Signal Processing Conference (EUSIPCO), 2016.

[5] YANG J C, JIANG B, LI B H, et al. A Fast Image Retrieval Method Designed for Network Big Data [J]. IEEE Transactions on Industrial Informatics, 2017, 13(5): 2350-2359.

[6] CHEN Z, WEN Y H, CAO J W, et al. A Survey of Bitmap Index Compression Algorithms for Big Data [J]. Tsinghua Science and Technology, 2015, 20(1): 100-115.

[7] 彭晏飞，张 维，訾玲玲，等. 一种具有双层信息损失优化结构的遥感图像检索方法 [J]. 计算机应用研究，2018，35(6)：1853-1857，1862.

[8] WANG S J, YAN S C, YANG J, et al. A General Exponential Framework for Dimensionality Reduction [J]. IEEE Transactions on Image Processing, 2014, 23(2): 920-930.

[9] BOUTSIDIS C, ZOUZIAS A, MAHONEY M W, et al. Randomized Dimensionality Reduction for <inline-Formula> <tex-Math Notation= "LaTeX" > k </tex-math></inline-formula>-Means Clustering [J]. IEEE Transactions on Information Theory, 2015, 61(2): 1045-1062.

[10] YANG J C, JIANG B, LI B H, et al. A Fast Image Retrieval Method Designed for Network Big Data [J]. IEEE Transactions on Industrial Informatics, 2017, 13(5): 2350-2359.

[11] HE T, WEI Y, LIU Z J, et al. Content Based Image Retrieval Method Based on SIFT Feature [C]. Xiamen: International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2018.

[12] WANG X L, WANG X, HOU A N. A Fast Quantization Tree Based Image Retrieval Method [C]//Proceedings on the International Conference on Artificial Intelligence (ICAI). Cairo: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016.

On Large Scale Image Indexing Technology Based on PCA-Binary Tree

ZHOU Xue-mei, PAN Duo

Department of Information Engineering, Sichuan Technology and Business College, Dujiangyan Sichuan 611830, China

Abstract: In order to solve the problem of dimension disaster in the image index in large data database, a large scale image retrieval technology based on cloud has been proposed in this paper. In the method, principal component analysis and binary trees have innovatively been introduced into image retrieval technology. First, scale invariant feature transform and speeded up robust features descriptor are used as the frame features. In the face of large-scale dimension features, the principal component analysis method is used to reduce the dimension of the frame feature, and a binary tree is used to represent the features after the dimension reduction to accelerate the research phase and reduce the storage space. Finally, image retrieval is realized. Experiments show that under the condition of reducing the dimension by 70%, the PR value of this method can reach the PR value under the traditional method of 20% dimensionality reduction. Compared with normal search, the search speed of this method is increased by 30%～50%.

Key words: big data; large scale image indexing; principal component analysis; binary trees; scale invariant feature transform

责任编辑 夏 娟