基于矢量量化的快速图像检索*

叶航军+, 徐光祐

(清华大学 计算机科学与技术系,北京 100084)

Fast Image Search Using Vector Quantization

YE Hang-Jun+, XU Guang-You

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: +86-10-62784141, E-mail: yehangjun98@mails.tsinghua.edu.cn, http://media.cs.tsinghua.edu.cn/

Received 2003-03-04; Accepted 2003-09-05

Ye HJ, Xu GY. Fast image search using vector quantization. *Journal of Software*, 2004,15(5):712~719.
http://www.jos.org.cn/1000-9825/15/712.htm

Abstract: Traditional indexing methods face the difficulty of 'curse of dimensionality' at high dimensionality. Accurate estimate of data distribution and efficient partition of data space are the key problems in high-dimensional indexing schemes. In this paper, a novel indexing method using vector quantization is proposed. It assumes a Gaussian mixture distribution which fits real-world image data reasonably well. After estimating this distribution through EM (expectation-maximization) method, this approach trains the optimized vector quantizers to partition the data space, which will gain from the dependency of dimensions and achieve more accurate vector approximation and less quantization distortion. Experiments on a large real-world dataset show a remarkable reduction of I/O overhead of the vector accesses which dominate the query time in the exact NN (nearest neighbor) searches. They also show an improvement on the indexing performance compared with the existing indexing schemes.

Key words: CBIR (content-based image retrieval); *k*-NN (nearest neighbor) search; high-dimensional indexing; curse of dimensionality; VQ (vector quantization); EM (expectation-maximization)

摘 要: 传统索引方法对高维数据存在"维数灾难"的困难.而对数据分布的精确描述及对数据空间的有效划分是高维索引机制中的关键问题.提出一种基于矢量量化的索引方法.该方法使用高斯混合模型描述数据的整体分布,并训练优化的矢量量化器划分数据空间.高斯混合模型能更好地描述真实图像库的数据分布;而矢量量化的划分方法可以充分利用维之间的统计相关性,能够对数据向量构造出更加精确的近似表示,从而提高索引结构的过滤效率并减少需要访问的数据向量.在大容量真实图像库上的实验表明,该方法显著减少了支配检索时间的 I/O 开销,提高了索引性能.

关键词: 基于内容的图像检索;*k*-近邻搜索;高维索引;维数灾难;矢量量化;期望最大化

中图法分类号: TP311 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant No.60273005 (国家自然科学基金)

作者简介: 叶航军(1976－),男,河南商水人,博士生,主要研究领域为基于内容的图像检索,高维数据索引机制,相关反馈方法;徐光祐(1940－),男,教授,博士生导师,主要研究领域为移动机器人,人机交互,基于内容的图像与视频检索,普适计算.

© 中国科学院软件研究所 http://www.jos.org.cn

基于内容的图像检索(content-based image retrieval,简称 CBIR)在过去十几年已经成为计算机视觉的重要研究领域[1].CBIR 系统通常用高维特征向量来表示图像,并使用某种距离度量来获得特征向量之间的相似度.通常的图像检索是一个相似度查询的过程,即利用相似度度量来寻找和查询最相似的图像.以特征向量来表示,图像库上的相似度检索就是寻找离查询图像的特征向量最近的前 k 个特征向量,也就是 k-最近邻(nearest neighbor,简称 NN)搜索问题.

k-NN 搜索问题可以通过对整个数据库的顺序查找来求解.特征向量按照它们到查询向量的距离做升序排列,前 k 个向量就是 k-NN 搜索的解.如果图像库里只有几千张图像,这个方法还是可行的.但随着图像库的迅速增大,简单的顺序扫描就因为计算量过大而变得不可行.对于使用相关反馈(relevance feedback,简称 RF)[2]技术的交互式图像检索系统而言,用户需要在每轮反馈后得到中间结果,所以每轮相似度检索的响应时间是 CBIR 系统成功的关键.这时就需要某种索引机制来加速相似度检索.

高维索引机制多年来都是数据库和多媒体领域的重要研究课题.研究者已经提出了很多基于空间划分和数据划分的索引方法来加速 k-NN 搜索,例如 gridfile[3],K-D-B 树[4],R*树[5],SR 树[6]等.这些传统方法在低维情况下(10 维以下)性能良好.然而一些研究[7,8]表明,这些传统方法在特征维数足够高的情况下(超过几十维),其性能还不如最原始的整个数据库顺序查找(强力搜索),这就是所谓的"维数灾难".

基于向量近似(vector approximation,简称 VA)的索引方法是目前在高维情况下惟一能优于顺序查找的一类精确索引方法.现有的基于向量近似的索引方法对图像库中的数据分布采用了比较简单的假设(各维独立的分布、单分量高斯分布等),用边缘概率分布来描述数据的整体分布,并使用标量量化的方法划分数据空间.过于简单的分布假设不能很好地描述真实图像库的数据分布;而标量量化的划分方法则无法充分利用维之间的相关性,导致产生的近似向量有较大的误差.这些缺陷严重影响了索引结构的性能.

本文针对大规模图像库的检索,提出了一种基于矢量量化(vector quantization,简称 VQ)的索引方法.该方法引入了性能更好的矢量量化器来代替简单的标量量化.它假设数据集具有混和高斯分布,并通过期望最大化(expectation-maximization,简称 EM)方法来估计分布参数.理论上,混和高斯分布可以表达任意的分布[9].因此,该方法不仅能处理均匀分布和高斯分布这些理想数据分布,还能很好地处理任意真实数据分布.

本文第 1 节介绍一些相关工作.第 2 节详细介绍基于矢量量化的索引算法.第 3 节给出实验结果并与其他方法作比较.第 4 节是结论与展望.

1 相关工作

近来研究者针对维数灾难提出了几种新的索引方法.其中基于向量近似(vector approximation,简称 VA)的索引方法是惟一能在精确 k-NN 搜索问题上优于顺序查找的方法.其他方法都是针对近似 k-NN 搜索问题提出来的.精确 k-NN 搜索是指能够正确得到 k-NN 结果集的检索方法,而在近似 k-NN 搜索返回的结果中,可能有些数据不属于正确的 k-NN 结果集,但一般来说,这些数据到查询向量的距离也比较接近正确的结果.

Weber 等人[7]提出的 VA 文件(VA-file)方法是第一个基于 VA 的索引方法.该方法把数据空间划分成 2^b 个超方体形状的胞腔(cell),其中 b 是用户指定的位串长度,用来近似估计原始特征向量.数据空间的每维分配 b_i 位,并被均匀地分割成 2^{b_i} 个区间,其中 $\sum_{i=1}^{d} b_i = b$, d 表示数据空间的维数.VA-file 方法并没有像 gridfile 或者 R 树那样把这些胞腔组织成分层结构(例如树),而是为每一个胞腔分配一个长度为 b 的惟一位串,并用这个位串近似表示落在该胞腔内部的原始特征向量.而 VA-file 本身就是一个由特征向量的紧凑几何表示组成的数组.如图 1(a)所示是一个二维数据集的 VA-file 划分.数据空间的每维都分配 2 位,并被均匀地分割成 4 个部分.

VA-file 上的 k-NN 搜索有两个阶段.第 1 个阶段是过滤阶段.首先顺序访问整个 VA-file,并根据特征向量的近似位串计算出该特征向量到查询向量距离的上、下界.如果某个向量的距离下界超过了目前为止遇到的第 k 小的距离上界,那么这个向量就可以被过滤掉,因为目前已经找到了至少 k 个更好的候选向量;否则,它就成为候选向量而进入下一个阶段.第 2 个阶段需要访问原始特征向量.经过第 1 阶段过滤后,只有很少一部分特征向量成为候选向量.这些候选向量按照它们到查询向量的距离下界做升序访问,并分别计算出到查询向量的精确距

© 中国科学院软件研究所 http://www.jos.org.cn

离.并不是所有的候选向量都需要访问,一般来说也只有很少一部分候选向量被访问.如果某个候选向量的距离下界超过了目前的第 k 个最近距离,整个查询过程就结束,当前的 k-最近邻就是查询结果.可以很容易地证明这个方法得到的查询结果就是精确 k-NN 搜索问题的解.

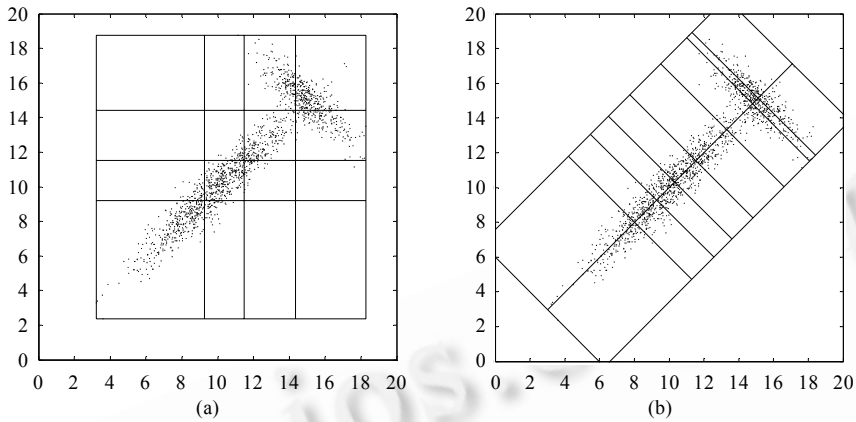

Fig.1 VA-file partition and VA$^+$-file partition of a 2-d dataset

图 1 一个二维数据集的 VA-file 划分和 VA$^+$-file 划分

VA-file 的性能取决于过滤的效率,而后者又依赖于在第 1 阶段所计算的距离上、下界的精度.上、下界的精度越高就能过滤掉更多不必要访问的向量,从而提高检索速度.特别需要指出的是,下界的精度直接影响在第 2 阶段需要访问原始向量的个数.所以更紧的下界可以直接减少最影响查询速度的随机磁盘访问次数.后来提出的几个基于 VA 的索引方法都是原始 VA-file 方法的改进,目的是要得到更为精确的上、下界.

VA-file 方法假设数据分布是均匀的且各维独立,因此独立地对各维进行划分.每次划分都等价于施加于该维上的一个标量量化器.事实上,VA-file 方法是最简单的矢量量化——分裂 VQ[10,11],即对各维依次独立进行标量量化,标量量化结果的笛卡尔积就作为矢量量化的结果.分裂 VQ 其实就是标量量化在矢量情况下的简单扩展,并没有利用各维之间的统计相关性,本质上还是标量量化.分裂 VQ 主要适用于各维独立或者接近独立的分布.如果数据各维有很强的相关性,分裂 VQ 就会有比较大的量化误差.从量化误差的角度来讲,矢量量化总是优于标量量化,并能从各维之间的统计相关性获益[12].对基于 VA 的索引方法而言,更小的量化误差就意味着更紧的上、下界,而后者则意味着更好的检索性能.

Wu 等人[13]针对数据分布的不均匀性,在各维上分别用高斯混合模型(Gaussian mixture model,简称 GMM)拟合数据的边缘概率分布,然后根据模型参数对数据各维进行独立划分.高斯混合模型对边缘概率分布的描述要比 VA-file 方法中的均匀分布假设更为精确,因此可以得到更精确的估计向量.用边缘概率分布来描述数据分布还是隐含了各维独立的假设,从而无法利用各维之间的相关性.

Ferhatosmanoglu 等人[14]提出了 VA$^+$文件(VA$^+$-file)的方法来处理非均匀分布的数据集.VA$^+$-file 方法首先使用 Karhunen-Loeve 变换(KLT)来去除各维之间的相关性,然后根据变换后各维的能量做不均匀的位数分配.在保证分配的总位数不变的前提下,方差(能量)大的维会得到更多的位数.VA$^+$-file 方法能够更精确地估计特征向量,在性能上也有了明显的提高.如图 1(b)所示为同一个二维数据集的 VA$^+$-file 划分.

VA$^+$-file 方法是原始 VA-file 方法的改进.但它本质上仍是一个分裂矢量量化器,除了它是在变换(KLT)域各维进行的标量量化以外,众所周知,KLT 只能消减各维之间的线性相关性,例如联合高斯分布.对于更一般的分布,各维之间的相关性一般不只是线性的.这时候,KLT 的作用就比较有限了.

2 基于矢量量化的索引算法

VA-file 的主要缺陷是过于简单的矢量量化方法.它假设数据各维是独立的,并独立进行标量量化.因此,它对各维独立分布的数据性能良好.而对于真实数据集的情况,直接在原始数据空间各维上的划分没有考虑各维

© 中国科学院软件研究所 http://www.jos.org.cn

之间的相关性,使得对数据空间的整体划分不佳.用这样生成的胞腔来近似原始向量,其误差显然很大,使得对上、下界的估计不够紧.Wu 等人的方法与 VA-file 方法类似,仍然使用了分布各维独立的假设,其主要改进是在对各维独立进行的划分方法上.

VA$^+$-file 方法在性能上比 VA-file 方法有了明显的提高,原因在于它合理地假设了数据各维之间存在相关性.VA$^+$-file 方法使用 KLT 去除各维之间的相关性,并对变换域的各维进行独立量化.KLT 能够较好地消除各维之间的相关性,但它把整个数据集看成一个类.实际上,KLT 是假设数据具有高斯分布.但如果数据来自多个类,这个假设就不能很好地揭示各维之间的内在相关性.因此,VA$^+$-file 方法适合于具有单一高斯分布的数据集.但对实际的图像库而言,图像一般来自不同的类别.这些图像的特征向量在数据空间上一般会聚成多个类.而对整个数据集进行 KLT 则使得特征向量的估计不够精确.

对向量精确估计的关键在于得到正确的数据分布,因此精确的数据分布模型会获得良好的索引效能.本文对数据分布采用了一个更合理的假设——数据集是由多个类别的混合模型产生的,并用高斯混合模型来表示.概率密度函数为

$$f(\boldsymbol{x})=\sum_{i=1}^{K}p_iG\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_i,\Sigma_i\right),\sum_{i=1}^{K}p_i=1,p_i\geq 0 \tag{1}$$

其中 K 是高斯中心的个数,p_i 是第 i 个高斯分量 C_i 的先验概率,$G\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_i,\Sigma_i\right)$ 是 C_i 的概率密度函数:

$$G\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_i,\Sigma_i\right)=\frac{\exp\left(-(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)/2\right)}{\sqrt{(2\pi)^d|\Sigma_i|}} \tag{2}$$

其中 $\boldsymbol{\mu}_i$ 是均值向量,Σ_i 是协方差矩阵.

混和高斯模型的几个性质使得它能很好地适应实际数据集的 k-NN 搜索问题:

1. 表达能力.理论上,混和高斯模型可以表示任意的分布[9].
2. 有效求解.存在有效的方法来估计混和高斯模型的参数,例如 K-Means[15]和 EM[16].
3. 适合聚类.每个高斯分量可以看作一个类.
4. 适合索引.对每个高斯分量可以通过 KLT 很好地消除各维之间的相关性.

高斯分量的总数 K 是由用户指定的.除此之外,每个高斯分量的权重(先验概率)p_i、均值向量 $\boldsymbol{\mu}_i$、协方差矩阵 Σ_i 都需要在整个数据集上进行拟合来确定.EM 方法类似于最大似然估计(maximum-likelihood,简称 ML),也是一种通用参数估计方法.它适用于含有隐含数据不能被直接观测的情况.对混和高斯模型而言,向量 $\boldsymbol{x}$ 对类的隶属关系是模糊的,需要用一个概率值来表示.这些概率形式的隶属度就是不能被直接观测的隐含数据.它们由各个高斯分量的参数来决定:

$$f_i(\boldsymbol{x})=\frac{p_iG\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_i,\Sigma_i\right)}{\sum_{h=1}^{K}p_hG\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_h,\Sigma_h\right)} \tag{3}$$

其中 $f_i(\boldsymbol{x})$ 就是向量 $\boldsymbol{x}$ 隶属于 C_i 的概率值.

如果用 $\boldsymbol{\Theta}=\{(p_i,\boldsymbol{\mu}_i,\Sigma_i),i=1,...,K\}$ 来表示混和高斯模型的参数集,那么数据集 D 的 log 似然(likelihood)函数就是

$$L(\boldsymbol{\Theta})=\sum_{\boldsymbol{x}\in D}\log\left(\sum_{i=1}^{K}p_iG\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_i,\Sigma_i\right)\right) \tag{4}$$

EM 算法首先使用某种方法(例如 K-Means 聚类)初始化对参数集 $\boldsymbol{\Theta}$ 的估计,然后用一个两步骤的过程用迭代方式更新对参数的估计:

1. 期望步(E-step).对数据集的每个向量 $\boldsymbol{x}\in D$,计算向量 $\boldsymbol{x}$ 隶属于 C_i 的概率值:

$$f_i^j(\boldsymbol{x})=\frac{p_i^jG\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_i^j,\Sigma_i^j\right)}{\sum_{h=1}^{K}p_i^jG\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_i^j,\Sigma_i^j\right)} \tag{5}$$

© 中国科学院软件研究所 http://www.jos.org.cn

2. 最大化步(M-step).更新混和模型参数:

$$p_i^{j+1}=\frac{\sum_{x\in D}f_i^j(x)}{N},\ \mu_i^{j+1}=\frac{\sum_{x\in D}f_i^j(x)\cdot x}{\sum_{x\in D}f_i^j(x)},\ \Sigma_i^{j+1}=\frac{\sum_{x\in D}f_i^j(x)\left(x-\mu_i^{j+1}\right)\left(x-\mu_i^{j+1}\right)^T}{\sum_{x\in D}f_i^j(x)} \tag{6}$$

其中上标 j 表示第 j 次迭代.当 log 似然函数 $L(\Theta)$ 的变化足够小的时候,整个迭代算法就可以退出.

VA-file 和 VA$^+$-file 方法都使用标量量化器来划分数据空间,无法利用各维之间的内在相关性,并导致了较大的量化误差及较松的距离上、下界.因此,本文引入了一种更通用的矢量量化方法来划分数据空间,以期获取更小的量化误差和更精确的向量估计.

通过 EM 算法估计出的混和高斯模型来构造数据集的索引.首先利用一个最优决策准则,把数据向量分到适当的类中:

$$g_i=\log p_i-\frac{\log|\Sigma_i|+(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}{2} \tag{7}$$

向量 x 最后被分到能最大化式(7)的类 C_h 中.实际上,它就是最小错误率的贝叶斯决策准则.

KLT 很适合具有单一高斯分布的数据集.因此,KLT 用来分别消除各个高斯分量中各维之间的相关性.下面是整个索引方法的算法流程:

1. 用 EM 算法为图像库拟合出混和高斯模型.
2. 用最小错误率的贝叶斯决策准则把每个向量分到适当的高斯类中.
3. 用 KLT 消除每个高斯分量中各维之间的相关性.
4. 根据各维的方差不均匀地分配位数.分配过程对各个类是独立进行的.
5. 对每个类依次量化,并为每个特征向量生成估计向量.

从矢量量化的角度而言,上面的算法流程是一个分类分裂矢量量化器.分类的目的是为了将来自不同类别的图像分开,使得类内各维之间具有更强、更明确的统计相关性,并使得矢量量化器的构造更为简单.根据矢量量化的理论,矢量量化器总会从更强的统计相关性获益.如图 2 所示为在同一个二维数据集上经过 EM 聚类后的划分.在该例中,数据点被聚成两类.在基于矢量量化的索引中,k-NN 搜索过程与在 VA-file 中的搜索过程类似,除了查询向量首先需要变换到每个类的 KLT 域中以外.

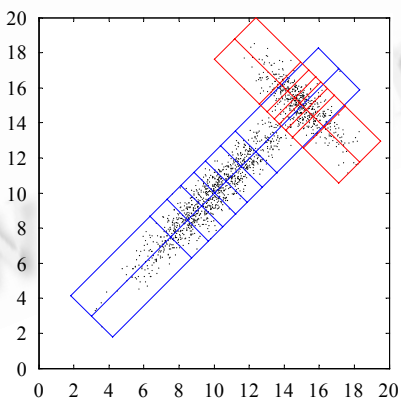

Fig.2 Partitioning after EM clustering (2 clusters totally)
图 2 EM 聚类后的划分(共有两个聚类)

3 实验结果

测试数据集是一个航拍图像库[17],有 275 465 幅图像,使用 60 维的 Gabor 纹理[18]作为图像特征.该数据集就是由 275 465 个 60 维向量组成的一个集合.这样大规模的数据集对任何高维索引方法而言都是一个挑战.现有的高维索引方法一般也选用该数据集来做性能测试.

© 中国科学院软件研究所 http://www.jos.org.cn

本文分别实现了 VA-file(VA),VA$^+$-file(VA$^+$),并用本文提出的基于 VQ 的方法来做性能测试.位串长度共实现了 1 位/维、2 位/维和 3 位/维这 3 种方案.注意,这个长度是每维的平均长度,具体的长度是根据每维的统计特性确定的.对于基于 VQ 的方法,用户需要指定高斯分量的个数.本文分别对 5 类(K=5,VQ-5)和 10 类(K=10,VQ-10)两种情况做了性能测试.从测试集里随机选取了 1 000 个向量作为 k-NN 搜索的查询向量.最后的测试结果是这 1 000 个查询向量性能数据的平均值.

我们使用的性能评价标准是两个阶段的过滤效率,即第 1 阶段后剩余的候选向量和第 2 阶段中访问的原始向量.表 1、图 3~图 5 是 3 种方法的测试结果(分别对 10-NN,50-NN 和 250-NN 的检索性能做了测试).

Table 1 Filtering efficiency for k-NN searches (in percentage)

表 1 对 k-NN 搜索的过滤效率(百分比)

		Filtering efficiency of first phase			Accumulated filtering efficiency of two phases		
		1bit	2bit	3bit	1bit	2bit	3bit
10 NN	VA	98.96	60.27	17.51	27.62	6.876	1.074
	VA$^+$	99.99	95.89	41.75	3.452	0.986 5	0.231 2
	VQ-5	37.82	10.96	0.859 7	0.946 6	0.177 5	0.042 7
	VQ-10	28.15	6.205	0.493 0	0.768 6	0.135 4	0.031 9
50 NN	VA	98.96	63.02	20.56	32.95	9.275	1.858
	VA$^+$	99.99	96.36	44.54	5.050	1.717	0.490 8
	VQ-5	40.92	13.94	1.599	1.719	0.416 3	0.126 2
	VQ-10	32.09	8.582	0.996 1	1.453	0.338 4	0.101 8
250 NN	VA	98.96	66.19	24.65	39.21	13.29	3.391
	VA$^+$	99.99	97.14	47.93	7.561	3.017	1.085
	VQ-5	44.84	17.90	2.980	3.241	1.013	0.389 7
	VQ-10	37.19	11.92	2.001	2.822	0.861 6	0.332 3

在第 1 阶段,基于 VQ 的方法在过滤效率上明显优于现有的 VA-file 方法和 VA$^+$-file 方法.而 VA$^+$-file 方法在第 1 阶段过滤性能不佳,在位数较少的情况下(1 位/维和 2 位/维)基本上没有过滤能力.尤其值得注意的是,基于 VQ 的方法(VQ-10)在第 1 阶段的过滤效率已经超过了 VA-file 方法两阶段的累计过滤效率.

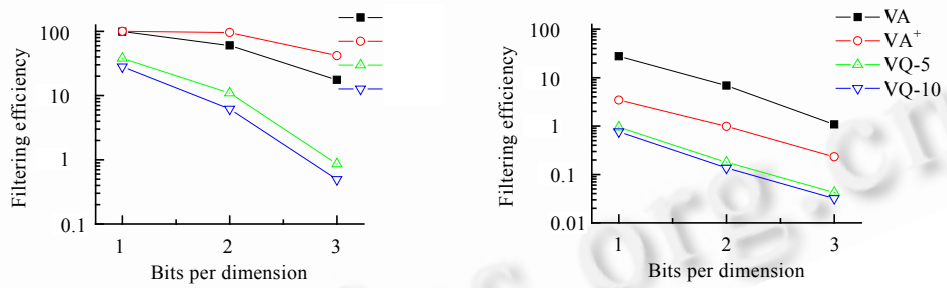

Fig.3 Filtering efficiencies of two phases for 10-NN searches (in percentage)

图 3 10-NN 搜索下的两阶段过滤效率(百分比)

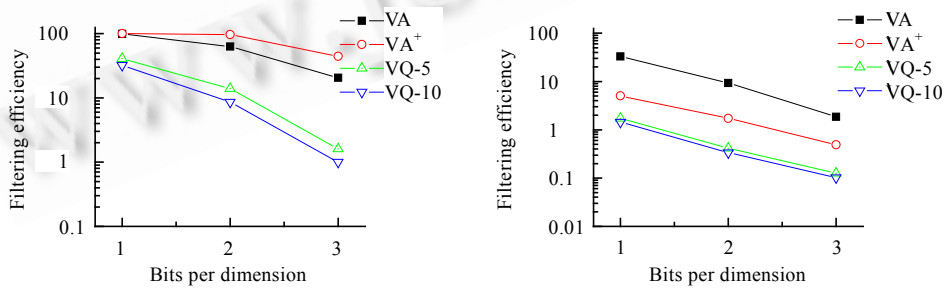

Fig.4 Filtering efficiencies of two phases for 50-NN searches (in percentage)

图 4 50-NN 搜索下的两阶段过滤效率(百分比)

© 中国科学院软件研究所 http://www.jos.org.cn

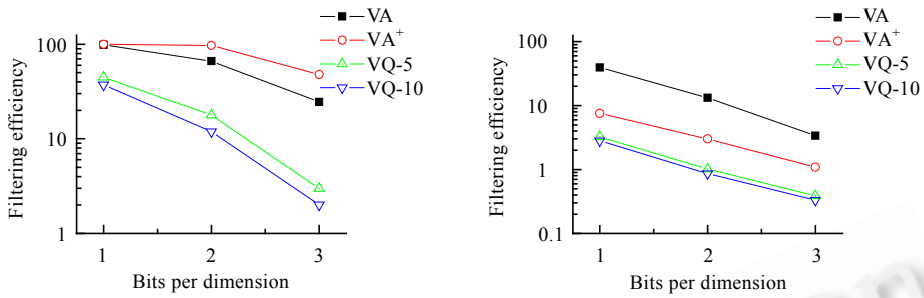

Fig.5 Filtering efficiencies of two phases for 250-NN searches (in percentage)
图 5 250-NN 搜索下的两阶段过滤效率(百分比)

在第 2 阶段中访问的原始向量正比于检索过程中的随机磁盘访问次数,而后者基本上支配了查询时间.基于 VQ 的方法(VQ-10)的两阶段累计过滤性能与其他 VA 方法相比有了显著的提高,过滤效率比 VA+-file 方法提高了 2~6 倍.虽然第 1 阶段的过滤性能不佳,但 VA+-file 方法在两阶段的累计过滤效率却明显优于 VA-file 方法,有 2~7 倍的提高.实际上,基于 VQ 的方法在 1 位/维的性能已经超过了 VA+-file 方法在 2 位/维的性能,而后者则超过了 VA-file 方法在 3 位/维的性能.基于 VQ 的方法只需要 VA+-file 方法一半大小的索引结构或者 VA-file 方法 1/3 大小的索引结构就能达到相同的检索性能.

基于 VQ 的方法在 10 类下的性能略优于在 5 类下的性能.如果类别再增加,应该会得到更高的性能.那么,为何不设定一个非常大的类别数呢?首先,太多类别下的参数估计非常困难,也很难做到精确.其次,每个类的量化器都需要自己的一份码书(codebook).在类别不太多的情况下,码书的开销和估计向量的总尺寸相比之下可以忽略不计.而对于类别太多的情况,则必须考虑码书的开销,而且这部分开销会明显降低索引的性能.另外,将查询向量变换到各类的 KLT 域也需要一部分时间上的开销,在类别过多的情况下,这部分开销也不能忽略不计.

4 结 论

本文针对大规模图像库的检索,提出了一种基于矢量量化的索引算法.现有的 VA 方法(VA-file 方法、VA+-file 方法等)假定图像库中的数据来自一个类,并使用标量量化器独立划分数据空间的各维.单类别假设使得对数据分布的估计不够准确.而过于简单的量化方法又导致了过大的量化误差.这些缺陷影响了相似度检索的检索效率.本文提出的方法利用对数据整体分布的直接估计(高斯混合模型)来代替现有精确索引方法中用边缘概率分布表示整体分布的方法,并引入了矢量量化器代替简单的标量量化方法来划分数据空间.从矢量量化的角度,本文引入了一种更通用的矢量量化器——分类分裂矢量量化器.它比简单的标量量化(分裂矢量量化)在性能上有明显的提高.基于 VQ 的方法在理论上更适合实际的数据集.对比实验的结果也证明,基于矢量量化的方法在 *k*-NN 搜索的性能上比现有的 VA 方法有显著的提高.

总之,基于矢量量化的方法是高维情况下精确 *k*-NN 搜索的有效索引方法.下一步的研究工作主要是寻找全局最优的矢量量化器以及索引机制和相关反馈方法的结合.

References:

[1] Rui Y, Huang TS, Chang SF. Image retrieval: Current techniques, promising directions and open issues. Journal of Visual Communication and Image Representation, 1999,10(4):39~62.

[2] Rui Y, Huang TS, Ortega M, Mehrotra S. Relevance feedback: A power tool for interactive content-based image retrieval. IEEE Trans. on Circuits and Systems for Video Technology, 1998,8(5):644~655.

[3] Nievergelt J, Hinterberger H, Sevcik K. The gridfile: An adaptable symmetric multikey file structure. ACM Trans. on Database Systems, 1984,9(1):38~71.

[4] Robinson J. The *k-d-b*-tree: A search structure for large multidimensional dynamic indexes. In: Edmund YL, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 1981. 10~18.

© 中国科学院软件研究所 http://www.jos.org.cn

[5] Beckmann N, Kriegel HP, Schneider R, Seeger B. The R*-tree: An efficient and robust access method for points and rectangles. In: Garcia-Molina H, Jagadish HV, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 1990. 322~331.

[6] Katayama N, Satoh S. The SR-tree: An index structure for high-dimensional nearest neighbor queries. In: Peckham J, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 1997. 369~380.

[7] Weber R, Schek H, Blott S. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: Gupta A, Shmueli O, Widom J, eds. Proc. of the 24th ACM Int'l Conf. on Very Large Data Bases (VLDB'98). New York: Morgan Kaufmann Publishers, 1998. 194~205.

[8] Beyer K, Goldstein J, Ramakrishnan R. When is 'nearest neighbor' meaningful? In: Beeri C, Buneman P, eds. Proc. of the 7th ACM Int'l Conf. on Database Theory (ICDT'99). Lecture Notes in Computer Science 1540, Berlin: Springer-Verlag, 1999. 217~235.

[9] Scott DW. Multivariate Density Estimation. New York: John Wiley and Sons, 1992.

[10] Gersho A, Gray RM. Vector Quantization and Signal Compression. Boston: Kluwer Academic Press, 1992.

[11] Sun SH, Lu ZM. Vector Quantization Technology and Applications. Beijing: Science Press, 2002 (in Chinese).

[12] Lookabaugh TD, Gray RM. High-Resolution theory and the vector quantizer advantage. IEEE Trans. on Information Theory, 1989,35(5):1020~1033.

[13] Wu P, Manjunath B, Chandrasekaran S. An adaptive index structure for high-dimensional similarity search. In: Shum HY, Liao M, Chang SF, eds. Proc. of Advances in Multimedia Information Processing—PCM 2001, the 2nd IEEE Pacific Rim Conf. on Multimedia. Lecture Notes in Computer Science 2195, Berlin: Springer-Verlag, 2001. 71~77.

[14] Ferhatosmanoglu H, Tuncel E, Agrawal D. Vector approximation based indexing for non-uniform high dimensional data sets. In: Proc. of the ACM Int'l Conf. on Information and Knowledge Management (CIKM 2000). New York: ACM Press, 2000. 202~209.

[15] Forgy E. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. Biometrics, 1965,21(3):768.

[16] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 1977,39(1):1~38.

[17] Manjunath BS. Aerial photo image database. 2000. http://vision.ece.ucsb.edu/datasets/

[18] Manjunath BS, Ma WY. Texture features for browsing and retrieval of image data. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1996,18(8):837~842.

附中文参考文献:

[11] 孙圣和,陆哲明.矢量量化技术及应用.北京:科学出版社,2002.

© 中国科学院软件研究所 http://www.jos.org.cn