第32卷 第22期 计 算 机 工 程 2006年11月
Vol.32 No.22 Computer Engineering November 2006
·博士论文· 文章编号：1000—3428(2006)22—0020—03 文献标识码：A 中图分类号：TP391.41

一种面向大规模图像库的降维索引新方法

贺 玲，吴玲达，蔡益朝
(国防科技大学信息系统与管理学院，长沙 410073)

摘 要：针对图像的72维HSV颜色特征，提出了一种新的降维方法。该方法在降维的过程中充分保留了图像颜色的本征特性。在降维的基础上，建立了一个新的索引机制，并以此加速大规模图像库的基于内容检索的进程。实验证明，该方法是行之有效的。
关键词：降维；本征维；基于内容的图像检索

A New Dimension Reduction Index Method for Large Image Databases

HE Ling, WU Lingda, CAI Yichao
(Information System & Management Department, National University of Defense Technology, Changsha 410073)

【Abstract】Aiming at the 72-dimensional HSV color feature in content-based image retrieval(CBIR), this paper proposes a new dimension reduction idea. It preserves the intrinsic merits of the original image color feature. Based on dimension reduction, it puts forward a new indexing structure to improve the performance of content-based retrieval of large image databases. Experiments show that this method is effective.
【Key words】Dimension reduction; Intrinsic dimensionality; Content-based image retrieval(CBIR)

随着信息技术以及多媒体技术的发展，规模越来越大的多媒体数据库出现在众多的科研、应用领域，大规模的图像库是其中重要的一种。

为了有效地管理这些大规模的图像数据并从中准确、快速地检索出有用的信息，仅仅依靠顺序扫描已经满足不了用户的需求。于是，适用于这些数据库的索引机制的建立成为一个重要的问题。

在基于内容图像检索的应用中，检索所针对的数据通常是从原始图像中抽取出来的几十维甚至上百维的高维特征向量。在这些情况下，传统索引机制的性能会急剧下降，甚至不按顺序扫描，这就是所谓的维度灾难[1]。为了有效解决这一问题，建立高效的高维索引机制，提高CBIR的性能，降维是一个直观而且有效的手段。

顾名思义，降维就是通过把数据点映射到更低维的空间上以寻求数据的紧凑表示的一种技术。由于丢弃了某些维度的属性值，它不可避免会带来信息的丢失，那么，尽量多保留原始数据的主要和重要特性是降维应该解决的首要问题。因此降维的过程与目标数据的特性密切相关，而作为高维数据的一个重要内蕴特征，本征维很好地反映了数据的本质维度。若能在降维处理中尽可能多保留数据的本征维信息，就可以最大限度地提高降维的性能。以此为出发点，本文针对图像的72维HSV颜色特征，提出了一种新的降维方法，该方法不仅实现简单，而且在降维的过程中保留了图像的绝大部分重要的颜色信息。在此基础上，建立了一个新的、适用于大规模图像库的高维索引机制，以此加速大规模图像库的基于内的容检索。

1 相关工作

由于简单直观且易于实现，基于降维的索引是很多高维索引机制应用最为普遍的方法之一。降维问题的模型(S,M)可定义如下：

$S=\{x_i\}_{i=1}^N$是D维空间中的数据集合；

降维映射$M:S\to L$，$x\to y=M(x)$

称y为x的降维表示。其中，L是d维空间的一个子集，且有$d\ll D$。

1.1 现有降维方法总结分析

根据不同降维算法的基本思想，可以把它们分成4类，即基于数据低维投影的降维、基于神经网络的降维、基于数据间相似度的降维和基于分形的降维。其中，主成分分析法和投影寻踪法属于第1类方法；自动编码网络、自组织映射网络和生成建模则属于第2类的范畴；第3类方法中比较经典常用的有多维尺度法、随机邻居嵌入、Isomap、局部线性嵌入和拉普拉斯特征映射。

上述3种降维方法应用较为广泛，而基于分形的降维则是近年来才得到关注的一类方法[2]，它最大限度地保留了数据集的本征特性，本文的研究重点也在于此。这些方法的优缺点如表1所示。

从表1可以看出，不同的方法在不同的领域都得到了广泛的应用。其中基于分形的降维能抓住数据集的本质特征，因此其降维结果能准确反映数据的本质属性。不过它的不足之处在于对本征维数进行估计时，通常需要数据集中数据的个数不小于$10^{D/2}$(D为数据维数)。而本文针对图像的HSV颜色特征进行的本征维估计没有受这一条件的限制，因而具有更普遍的应用范围和应用前景。

基金项目：国家自然科学基金资助项目(60473117)
作者简介：贺 玲(1976－)，女，博士生，主研方向：多媒体信息系统，高维数据索引；吴玲达，教授、博导；蔡益朝，博士生
收稿日期：2005-11-24 **E-mail：**heling6159@163.com

表 1 降维方法优缺点比较

降维方法	优点	不足
主成分分析(PCA)	概念简单、计算方便，具有最优线性重构误差	不能处理非线性数据；没有明确的准则来确定应该保留多少主成分
核 PCA	PCA 在非线性应用领域的扩展	算法的性能极大地依赖于核的选择
主曲线	PCA 在非线性应用领域的扩展	算法的收敛性未得到证明
投影寻踪	可有效排除噪声数据的干扰	计算量大；不适用于高度非线性数据
自组织映射	能很好用于高维数据可视化	无法定义可以优化的代价函数；缺乏收敛性的一般定义
贝叶斯神经网络	能得到参数的后验分布	先验分布是任意选择的
生成拓扑映射	能很好地应用于数据可视化	不适用于硬降维
多维尺度	可很好地保持数据间的差异性	没有统一的准则来评价所得到的嵌入维的质量
局部线性嵌入	是一种探索性的数据分析方法，可很好揭示高维空间中的低维结构；计算简单直观	目前主要用于可视化，在其它方面尚未得到充分应用
基于分形的降维	抓住了数据集的本征特性；能得到数据的分数维估计	为了获得 D 维数据准确的本征维估计，数据集中数据的个数 $N \geq 10^{2/D}$

1.2 本征维及其相关概念

高维(D维)空间的样本数据，一般不可能弥漫于整个R^D空间，否则数据集中就不会有什么有用的信息。这些数据实际上处于一个高维空间的低维流形上，即一个降维的“曲面”上，该流形的维数即为数据的本征维数，而D只是数据的表象维数。换句话说，一个数据集的本征维是指该数据集所表示的空间对象的实际维度，而不管其所在的空间维数。例如三维欧式空间中的一条曲线的本征维数只是 1。从实用的角度来看，降维的出发点是在保留原结构信息的条件下尽可能降低维数，简化高维数据的表示，因而对于本征维数不一定要求非常精确的估计，能够显著降低维数并为后续工作提供便利，就是令人满意的选择。

需要注意的是，“本征维数估计”这个提法本身就不够严格，因为在几乎所有情况下，用以估计本征维数的观察数据都是高维空间中有限个离散的向量，估计总是依赖于不同的人对于各自的不同问题所使用的不同准则，诸如流形的光滑性、噪声的影响大小等。所以对于同样的数据点很可能得出不同的结果。如图 1 所示，可以用一个一维的流形(曲线)来拟和图中的黑点，可是它们也可能位于一个二维流形上。因此为了获得本征维数的令人满意的估计，就需要一个不断调整和检验的过程。

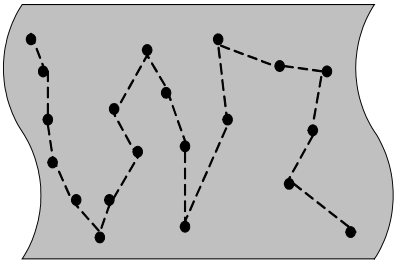

图 1 本征维数

2 基于本征维的降维及索引

如前所述，由于本征维决定了高维数据的本质内容和属性，因此用本征维来指导降维处理，可以尽可能完整地保留原始数据的有用信息。本节将详细介绍针对 HSV 颜色特征的降维处理，以及在此基础上的索引和近似搜索算法。

2.1 HSV 颜色模型

图像的物理特征可以通过多种方式来体现，颜色、形状和纹理等都是常用的颜色特征。图像索引机制的建立与图像特征的具体表现形式密切相关。本文则主要针对图像的 HSV 颜色特征，探讨合适的降维机制。

HSV 颜色模型是一种基于感知的颜色模型，它把彩色信号表示为 3 种属性：色调(Hue)，饱和度(Saturation)和亮度(Value)，这种颜色模型用 Munsell 三维空间坐标系统表示。其中 H 表示从一个物体反射过来的或透过物体的光的波长，一般通过颜色名称来进行辨别，并用角度-180°~180°度量。饱和度 S 指颜色的深浅，它用百分比来度量，变化范围从0%～100%。亮度 V 是颜色的明暗度，也用百分比度量。

一般来说，从图像中直接得到的特征值通常都是 RGB 值，因此还需要把 RGB 值转换为 HSV 值。其详细的转换过程可参照文献[3]，这里不再赘述。

得到了相应的h、s和v值之后，还应该对HSV空间进行适当的量化后再计算直方图，这样能大大减小所需的计算量。量化之后，即可将h、s和v这 3 个分量在一维矢量上(记作L)分布开来，并且L的取值范围可以确定为[0，1，… ，71]。这样就获得了 72 柄的一维直方图[3]。

2.2 HSV 特征的降维及对应的搜索算法

从上面对 HSV 颜色模型的介绍可知，HSV 特征的每一维都代表了某种颜色值在对应图像中所占的比例。因此对于 2 幅图像来讲，若某种颜色值在一幅图像中所占的比例为 30%、在另一幅图像中所占的比例为 25%，那么它们在这种颜色分量的相似度就应该为 25%，这可以简单地通过集合的交集特性来理解。基于此，2 幅图像之间的相似度计算通常采用直方图交的方法。设有 2 幅图像的 HSV 特征为

$$X=(Hsv_{x0},Hsv_{x1},...,Hsv_{x71})，\ Y=(Hsv_{y0},Hsv_{y1},...,Hsv_{y71})$$

那么采用直方图交的形式，X 所代表的图像与 Y 所代表的图像之间的相似度可表示为 $Sim(X,Y)=\sum_{i=0}^{71}\min(Hsv_{xi},Hsv_{yi})$。

对于上式来说，只有 2 幅图像的每个分量的值都尽可能大，才有可能使其相似度值尽可能大。但是，经过归一化后的 1 幅图像的所有 72 维 HSV 特征值的和等于 1。也就是说，如果某些维上的分量很大，必然有其它维上的分量很小。因此，对于一幅图像而言，如果某些维上的特征值之和超过 80%或是更多，也就是说一幅图像至少 80%的颜色是由这些维决定的，那么仅保留这些维就足以准确表示原始图像的颜色信息。依据这一思想进行降维，能够最大限度地保留原始数据的根本特性，不会造成主要信息的损失。

而所谓本征维，就是能代表一个数据本征特性的所有维度的集合。因此对于用 72 维 HSV 颜色特征刻画的图像数据来讲，能表示图像的主要颜色信息的那些维就可以称为本征维。而且在这一求解过程中，本征维是一个相对确定的维度集合，不会出现图 1 所示的模糊状况。

针对 HSV 颜色特征，基于本征维的降维可有如下简单的过程化描述：

Input：图像数据集的原始 HSV 颜色特征库，控制参数 k

Output：各图像本征维的 HSV 特征及相应的维度

Step1 若特征库非空，取出当前第 1 条记录，对 72 维特

征值从大到小排序；

Step2 在另一个数据表中记下排序后的前 1 到 k 维特征值及相应的维度，并转向原始特征集的下一个记录；

Step3 重复 Step 1 和 Step 2，直至原始特征集中的数据全部处理完毕。

通过上述步骤降维之后，72 维的特征向量在不损失本质信息的前提下降到了 k 维，这无疑能大大简化对特征数据库的基于内容相似检索。而且，在降维的过程中，已经形成了一个由本征维及这些维上的属性值组成的一个索引表。依据这一索引表，能够方便地实现基于内容的图像检索。

以基于原始数据库的示例检索为例，相似搜索算法可描述如下：

Input： 样本图像，过滤参数 F，相似度 Sim

Output：与样本图像相似度大于 Sim 的所有结果集 R

Step1 从索引表中得到样本图像的本征维信息；

Step2 根据样本图像的 1 到 k 个本征维，从原始特征集中找出在这些本征维上至少有一维的值大于样本图像的所有图像，将其加入第 1 阶段的结果集 S 中；

Step3 对于 S 中的所有图像，依次从索引表中得到其本征维信息，再将这些本征维上的属性值与样本图像的相同维的属性值进行比较，记下后者在这些维上取值大于前者的维数之和，记作 Filter；

Step4 若 Filter≥F，把相应的图像加入中间结果集 S’中；

Step5 对于 S’中所有的图像，在原始特征空间上依次计算其与样本图像的相似度，若该相似度值大于或等于 Sim，把相应的图像加入最终的结果集 R 中。

上述搜索算法中过滤参数 Filter 的作用就是使用户能同时考虑样本数据和搜索空间中其它数据的特性，即考虑局部最优性的同时也考虑整体最优性，从而保证检索结果的准确和完整。

3 实验

实验在一个含有 1 万多个样本数据的、由 SQL Server 组织起来的图像数据库上进行。通过对所有的样本统计分析可知，取前 10 个最大维值，它们的和平均达到了 86.179 0%，也就是说，取控制参数 k=10，把 72 维特征值降至 10 维，就足以描述原始图像数据。

实验测试了 Sim 取值分别为 50%、60%和 80%时的相似搜索情况。在对搜索算法的性能评估中，除了查询响应时间之外，算法的查到率和查准率是 2 个重要的评价指标。所谓查到率，是指查到的相关图像的数目与库中实际相关数目的比值；而查准率则是指查到的相关图像数目与查到图像数目的比值。

但是由于“相关”的概念是很模糊的，人们从肉眼上直观感觉到的“相关”很难让机器去精确识别，因此对于大规模的图像数据库而言，用户也很难找出与示例图像相关的所有图像的数目。

考虑到顺序查找是没有任何过滤处理的搜索过程，从而可以认为顺序查找得到的结果都是与样本图像相关的图像，因此顺序查找的查到率可以认为是 100%。那么本文基于降维处理的搜索算法的查到率就可以定义为：查到率＝查到的图像数/顺序查找的图像数目。为此，首先分别实现了取不同相似度时的顺序查找，针对本文给出的一个示例样本(在图像数据库中，样本 ID 为 3 208)，顺序查找查到的图像数目分别为 386、209 和 67。由于在这些情况下，顺序查找都计算了样本图像与所有图像之间的相似度大小，因此其查询时间都为 10.682s。

对于基于降维的搜索来讲，根据查询相似度的不同，试验中的过滤参数 Filter 依次取 4、6 以及 8，不同情况下的查询响应时间(Time)及查到率(Recall)如表 2 所示。

表 2 相似搜索中针对不同相似度的查询响应时间及查到率

	Time	Recall
Sim=50%	1.816s	88%
Sim=60%	1.225s	100%
Sim=80%	0.962s	91%

从表 2 可以看出，本文算法的查询相应时间远远小于了顺序查找的时间。不过，对于相似度阈值较小的查询，本文算法的性能还不能得到突出的体现。近似搜索的相似度阈值越大，利用本文算法进行查询时不仅响应时间更短，搜索性能也随之提高，这是因为 Filter 参数的取值随着相似度阈值的升高有了相应的调整，也就是说 Filter 的值越大，从 S 中过滤掉的可能无关的候选数据越多。但是，正是 Filter 的增大，当相似度阈值取 80%的时候，也过滤掉了一些相关图像。这说明，Filter 的取值也是影响算法性能的一个重要因素。

4 总结

针对图像的 72 维 HSV 颜色特征，本文提出了一种新的基于本征维的降维索引机制。结合本征维的概念，并基于 HSV 颜色模型的特定涵义，本文首先把决定图像主要颜色信息的那些维界定为图像特征的本征维，然后根据这些本征维对 72 维的 HSV 颜色特征进行降维处理，从而构建了一个高维索引机制。

为验证这一降维索引的有效性，文章针对图像的基于内容相似检索进行了实验。从实验中可以看出，本文算法不仅在查询响应时间上大大优于顺序查找的响应时间，而且其查到率也达到了很好的效果。

今后的工作将主要集中在通过反复的实验得到 Filter 因子的取值规律，以使得本文基于本征维的降维索引能更多地提高 CBIR 的性能。

参考文献

1 Chavez E, Navarro G. Probabilistic Proximity Search: Fighting the Curse of Dimensionality in Metric Spaces[J]. Information Processing Letters, 2003, 85(1): 39-46.

2 Camastra F, Vinciarelli A. Estimating the Intrinsic Dimension of Data with a Fractal-based Method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(10): 1404-1407.

3 李国辉, 柳 伟, 曹莉华. 一种基于颜色特征的图像检索方法[J]. 中国图形图像学报(A 版), 1999, 4(3): 248-255.