DOI:10.13878/j.cnki.jnuist.2017.06.006

周文罡[1]　李厚强[1]　田奇[2]

图像检索技术研究进展

摘要

近年来,互联网上视觉数据呈现出爆炸式的增长,越来越多的研究工作围绕图像搜索或图像检索技术而展开.早期的搜索技术仅采用文本信息,忽视了视觉内容作为排序的线索,导致搜索文本和视觉内容不一致.基于内容的图像检索(CBIR)技术充分利用视觉内容识别相关图像,在近几年来获得了广泛关注.在图像检索中,最根本的问题是意图鸿沟和语义鸿沟,围绕该问题,近年涌现出大量的基于内容的图像检索的技术.本文主要对2003—2016年间提出的相关图像检索方法进行总结、分类和评估,并对未来的潜在研究方向进行讨论.

关键词

图像检索;视觉表征;索引;相关性度量;空间上下文;检索重排序

中图分类号 TP391.41

文献标志码 A

收稿日期 2017-08-28

资助项目 国家自然科学基金(61472378)

作者简介

周文罡,男,博士,副教授,博士生导师,研究方向为多媒体信息检索和计算机视觉.zhwg@ustc.edu.cn

1 中国科学技术大学 信息科学技术学院,合肥,230026

2 美国德州大学圣安东尼奥分校 计算机系,San Antonio,TX,78249,USA

1 背景介绍

随着数码设备的普及以及网络技术的飞速发展,数十亿人在网上共享和浏览照片.图像检索(CBIR)致力于从大规模图像数据库中检索出与文本查询或视觉查询相关的视觉内容.自20世纪90年代以来,图像搜索引起了多媒体等领域研究人员的广泛关注[1].传统的图像搜索引擎通常基于图像周边围绕的元数据信息,例如标题和标签,来索引多媒体视觉信息.但是由于这些文本信息可能与视觉信息不一致,其检索结果可能不可靠.为避免这种问题,基于内容的图像检索技术被引入,并在近些年取得了很大的进步.在基于内容的图像搜索中有2个基本的挑战,分别是意图鸿沟和语义鸿沟.意图鸿沟指的是用户很难通过一个查询,例如一张图像或是一个素描图,精确地表达他所期望的视觉内容;语义鸿沟是指采用一个低阶的视觉特征来描述一个高阶的语义内容是很困难的[2-4].为了缩小这种鸿沟,学术界和工业界做出了大量的研究工作,并取得了长足进展.

从20世纪90年代初到21世纪初,很多基于内容的图像搜索的相关研究被发表,已有综述性论文讨论过这些研究[5-7].在21世纪初期,随着一些新的见解和方法的提出,CBIR向另一个研究趋势发展.尤其是2项开创性的研究工作为大规模多媒体库中基于内容的视觉检索的重大进展铺平了道路.第1个是局部视觉特征SIFT的提出[8].SIFT被证明具有极好的描述性和区分性,以捕获各种多媒体数据中的视觉内容.它具有对旋转和尺度变换的不变性,同时也对光照变化具有很好的鲁棒性.第2个工作是词袋模型(Bag-of-Visual-Words,BoW)的提出[9].当用于信息检索时,BoW模型通过量化图像中包含的局部视觉特征生成图像的紧凑表达.同时,BoW模型可以适应于倒排索引结构,可以更好地应用于大规模图像检索.

基于上述开创性的工作,最近10年中涌现出大量的基于多媒体内容的图像检索研究工作[10-29].然而,在工业界,一些基于内容的图像搜索引擎各有所侧重,例如Tineye(tineye.com)、Ditto(ditto.us.com)、Snap Fashion(www.snapfashion.co.uk)、ViSenze(www.visenze.com)、Cortica(www.cortica.com)等.Tineye于2008年5月推出了10亿幅反向图像搜索引擎.到了2017年1月,Tineye数据库中索引的图像已经到达了170亿幅.不同于Tineye,Ditto特别关注于商标图像,通过Ditto可以发掘社交媒体上共享的照片中的商标信息.

从技术上讲,基于内容的图像检索中存在 3 个关键问题:图像的表达、图像的组织和图像相似度度量.现有的方法可以基于这 3 个关键问题进行分类.

图像表达是基于内容的视觉检索的本质性基础问题.为了方便比较,一幅图像可以被转换到某种特征空间,以实现隐式的对齐,从而消除背景和潜在变形的影响,同时保持内在视觉内容的区分.事实上,如何进行图像表征是计算机视觉任务中的一个根本性问题.通常,一幅图像被表达成一个或多个视觉特征.这个表达须具有描述性和区分性,以便于区分相关与不相关的图像.更加重要的是,人们期望图像表达对各种变化(例如平移、旋转、缩放、光照变换等)具有不变性.

在多媒体检索中,视觉数据库通常非常巨大.一个非常重要的问题是如何组织数据库,以便于当给定一幅查询图像时,能够有效地识别出相关结果.受到信息检索的启发,许多现有的基于内容的视觉检索算法和系统利用经典的倒排索引结构索引大规模的视觉数据库.一些基于哈希的技术也以同样的视角被引入到索引中.为了实现这一目标,视觉码本学习和高维视觉特征的特征量化等技术被引入,嵌入空间上下文信息也可以进一步提高视觉表示的辨别能力.

理想情况下,图像间的相似度须反映语义上的相关性,然而因为语义鸿沟的存在使其变得困难.在基于内容的图像检索中,图像的相似度一般被定义为视觉特征的加权匹配结果.现存算法中图像相似度定义可以看成是不同的匹配核[30].

本文主要概述 2003 年至今的 10 多年间图像检索的研究工作.对于 2003 年以前的工作,建议读者阅读先前的综述论文[5-7].最近,也有一些关于 CBIR 的综述文章[2-3,31].文献[31]从数据库规模的角度总结了过去 20 年的图像搜索工作;文献[3]在社会图像标签的背景下,对最新的 CBIR 技术进行了回顾,重点论述了 3 个紧密联系的问题:图像标签分配、优化和基于标签的图像检索.本文则从不同的视角讨论了 CBIR,更多地强调通用框架方法方面的进展.

在后续的章节中,本文首先简要回顾基于内容的图像检索的通用框架,然后分别讨论这个框架中的 5 个关键模块;之后,介绍普遍使用的测试数据集和评估标准;最后,讨论未来潜在的发展方向并做总结.

2 通用流程图概述

基于内容的图像检索是多媒体领域的一个热点研究问题.图像检索的通用流程如图 1 所示.图 1 所示的视觉检索系统由离线和在线 2 个阶段组成.在离线阶段,通过图像爬虫工具构建图像数据库,将数据库中的每张图像表达成特征向量并构建索引.在线阶段包含 6 个模块:用户意图分析、查询构成、图像表达、图像相关度评分、搜索重排序和搜索结果浏览.图像表达模块在离线和在线阶段共享.本文不包含图像数据库爬取、用户意图分析[32]和检索结果浏览[33].这些方面的研究可以参考前人的工作[6,34].本文的后续部分集中讨论其他 5 个模块,即:查询构成、图像表达、数据库索引、图像相关度评分和检索重排序.在后面几节,本文总结每个模块的相关工作,讨论和分析每个模块在关键问题上所采取的策略.

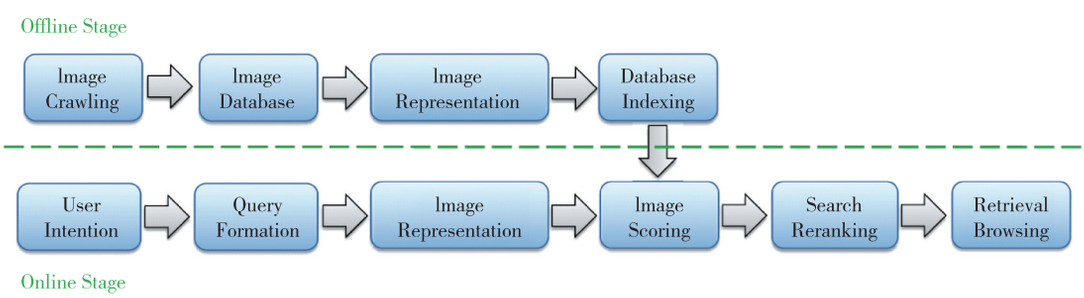

图 1 基于内容的图像检索的基本框架(绿线以上代表离线操作,绿线以下是在线操作,本文将详细介绍查询构成、图像表征、数据库索引、图像评分和检索重排序 5 个方面)

Fig. 1 The general framework of content-based image retrieval.(The modules above and below the green dashed line are in the off-line stage and on-line stage,respectively.In this paper,we focus the discussion on five components,i.e.,query formation,image representation,database indexing,image scoring,and search reranking)

3 查询构成

进行图像检索时,用户往往将自己的意图用具体的视觉查询表达出来.查询图像的质量对检索结果有显著的影响.一个良好的、明确的查询可以有效地降低检索的难度,获得更加满意的结果.大体上,根据图像的不同类型,查询构成方式可以分为以下几种:基于示例图像、基于草图、基于颜色图、基于上下文图等.如图 2 所示,不同的方法会导致明显不同的结果.在下文中,我们分别讨论这些代表性的查询构成方式.

最直观的查询构成方式是示例图像,用户使用一张查询图进行查询,希望检索到更多、更好的同一张图片或者具有相同语义的相似图片.例如:一幅图像的所有者可能需要了解他/她的图片是否未经允许而被某些网页使用;一名网警可能希望通过检查恐怖组织的 logo 是否出现在网络图片或视频中来反恐,等等.为了降低背景的影响,在检索时,可以框选出示例图片中的感兴趣区域.由于示例图片是客观的,不受人的主观影响,很容易通过对它做定量分析,从而优化对应的算法.因此,通过示例图片搜索是基于内容的图像检索系统中被研究最多的方法[9-10,35-36].

除了通过示例图片检索,用户也可以通过草图来表达意图[37-38].在这种方式中,查询是一张轮廓图.轮廓比较接近语义表达,它能够帮助系统从语义的角度检索到符合用户意图的结果[27].最初,基于草图的检索只能用于一些特定的图像,比如剪切画[39-40]和简单模式图片[41].一个里程碑式的工作是 Edgel 用草图搜索自然图像[42].草图也在一些搜索引擎中得到了应用,比如 Gazopa(www.gazopa.com)和 Retrievr(http://labs.systemone.at/retrievr/).然而,基于草图的搜索有 2 个不足之处.首先,除了太阳、鱼、花等可以被简单形状表达的对象,用户很难快速地通过轮廓表达出搜索目标;其次,由于数据库中的图像通常是自然图像,需要设计专门的算法将自然图像转化成与用户意图相符的轮廓图.

另一种查询构成形式是颜色图.这类系统会提供给用户一种格子状的调色板,用户利用它指定图像不同区域的颜色分布,从而检索具有相似颜色分布的图像[43].通过嵌入粗略的形状,颜色图允许用户和系统进行交互来提高检索效果.但这种方法受限于所能表达的潜在语义概念.另外,在图像采集时,颜色和亮度发生变化是很常见的,这将严重影响颜色特征的可靠性.

上述的查询构成方法便于用作输入,但仍然难

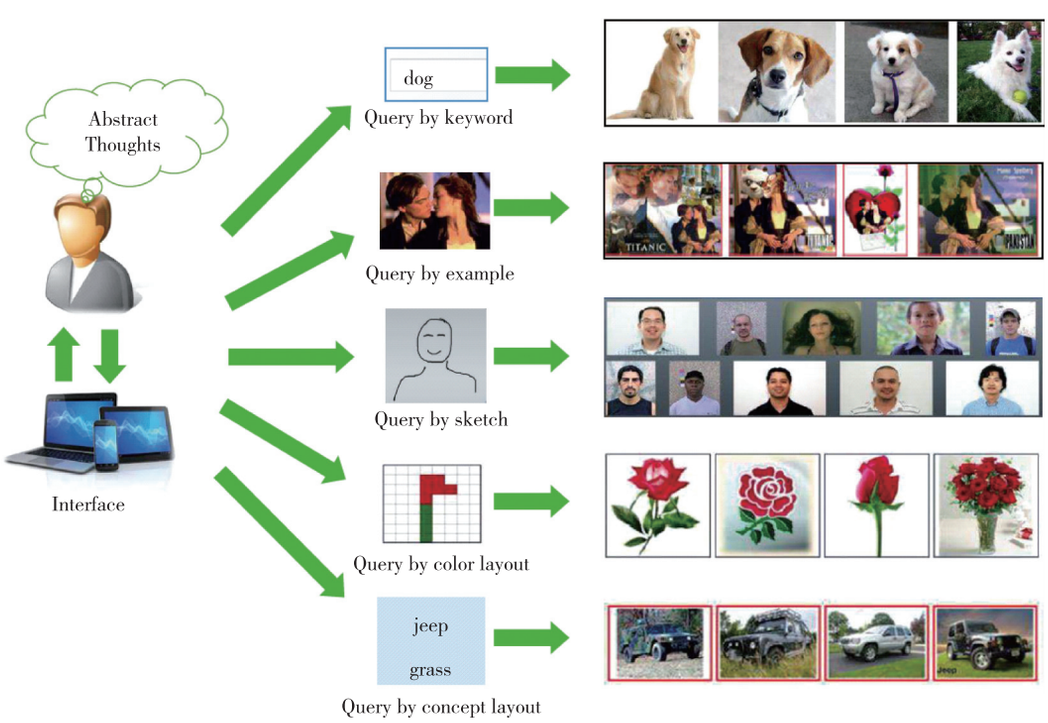

图 2 不同的查询形式及其检索结果

Fig. 2 Illustration of different query schemes with the corresponding retrieval results

以准确表达语义意图.为了解决这个问题,Xu等[44-45]提出在查询图像的某些位置加文字来表达语义概念.文献[46]基于排序SVM模型,也对这种结构化的目标查询进行了研究.这类查询要求数据库图像和查询图像中的目标或者场景被提前识别出来.

值得注意的是,上述几个目前被大多数工作采用的方法中,查询都是单张图像,这在某些情况下可能不足以反映用户的意图.如果提供更多图像作为查询,则可以使用一些新的策略来共同表达查询或者融合单一特征的检索结果[47].这或许是一个有意思的课题,尤其是在视频检索中,此时查询是一个时序的视频片段.

4 图像表达

在基于内容的图像检索中,关键问题是如何有效测量图像之间的相似性.由于视觉目标或场景可能经历不同的变换,直接在像素层面比较2幅图像是难以实行的.通常,首先从图像中抽取视觉特征,然后将其变换成固定长度的向量作为图像表达.考虑到大规模数据集与有效查询响应之间的矛盾,有必要整合视觉特征以加速索引和比较的过程.为了实现这个目标,使用视觉码本进行量化常被用于特征聚合编码过程.除此之外,作为视觉数据的一个重要特性,空间上下文信息对于提高视觉表达的区分性是至关重要的.

基于上述讨论,我们可以将2幅图像 和 之间的内容相似性在数学上形式化如下:

$$S(\quad)=\sum_{x\in}\sum_{y\in}k(x,y)= \tag{1}$$

$$\sum_{x\in}\sum_{y\in}\boldsymbol{\phi}(x)^{\mathrm{T}}\boldsymbol{\phi}(y)= \tag{2}$$

$$\boldsymbol{\Psi}(\quad)^{\mathrm{T}}\boldsymbol{\Psi}(\quad). \tag{3}$$

基于式(1),这里产生3个问题:

1)如何使用一组视觉特征$\{x_1,x_2,\cdots\}$描述图像内容 ?

2)如何将不同长度的特征集合 $=\{x_1,x_2,\cdots\}$变换成固定长度的特征向量$\boldsymbol{\Psi}(\quad)$?

3)如何有效计算2个固定长度的向量之间的相似性$\boldsymbol{\Psi}(\quad)^{\mathrm{T}}\boldsymbol{\Psi}(\quad)$?

上述3个问题分别对应于特征抽取、特征编码与聚合和数据库索引.特征编码与聚合过程包括视觉码本学习、空间上下文嵌入和量化.在本章节,我们讨论在图像表达,包括特征提取、视觉码本学习、空间上下文嵌入、量化和特征聚合这些关键问题上的相关工作.数据库建立索引将在下一章节中讨论.

4.1 特征提取

传统上,视觉特征被启发式地设计,并分为局部特征和全局特征.除了那些手工设计特征,近年来基于数据驱动的特征学习也获得极大发展.下面将分别讨论这2种特征.

4.1.1 手工设计的特征

在早期的基于内容的图像检索算法与系统中,全局特征通常将颜色[43,48]、形状[42,49-51]、纹理[52-53]和结构[54]转化为单一全局表达来描述图像内容.作为全局特征的一个重要代表,GIST特征[55]在生物学上具有合理性,其计算复杂度较低,已经被广泛用来评估近似最近邻搜索算法[56-59].由于表达紧致且计算高效,全局视觉特征适用于大规模数据库的图像拷贝检测[54],但对处理背景复杂的图像可能效果不佳.典型地,全局特征能够被用作局部视觉特征的补充部分以提高近复制图像检索的准确度[24].

自从Lowe首次提出SIFT特征[8,60]以来,局部特征已经被大量的基于内容的图像检索工作用作图像表达.通常,局部特征抽取包括2个重要阶段,即兴趣点检测和局部区域描述.在兴趣点检测中,一些特定尺度的关键点或区域被高可重复率地检测到.这里的重复率意味着兴趣点在不同的变换或改变中仍能被检测到.常用的检测子包括差分高斯[8]、最大稳定极值区域[61]、Hessian仿射检测子[62]、Harris-Hessian检测子[63]和FAST[64].在兴趣点检测中,可实现平移和尺度变化的不变性.与上述方法不同,不使用任何显式的检测子,仅通过均匀地、密集地采样图像平面获得兴趣点也是可能的[65].

在兴趣点检测之后,抽取一个或多个描述子用来描述以兴趣点为中心的局部区域的视觉外观[66].一般情况下,描述子被设计成具有旋转不变性,并且对仿射变换、噪声和光照变化等保持稳定的形式.除此之外,描述子也应该具有区分性,使得它可以从很多图像特征构成的大集合中以很高的概率获得正确的匹配.在很多大数据集视觉应用中,都特别强调这种性质.最常用的具有以上性质的描述子是SIFT特征[8].作为SIFT的变形,SURF可以获得与SIFT可比的性能,但计算更有效[67].

一些研究人员探索基于SIFT的提升或扩展.在文献[23]中,Arandjelovic等在原始的SIFT描述子上进行平方根-归一化操作获得root-SIFT特征.尽管操作简单,root-SIFT已经被证明可有效地提高图像

检索的精度,并能够稳定地应用于很多基于 SIFT 的图像检索算法中[68].Zhou 等[36]提出使用原始描述子的 2 个中值作为阈值生成 SIFT 描述子的二值化特征,获得的二值化 SIFT 产生一种新的图像检索的索引方法[69].Liu 等[70]扩展了二值化 SIFT,他们首先通过维度对比生成一个二值化比较矩阵,然后灵活地划分矩阵元素到各部分,每个部分被哈希到 1 bit.在文献[21]中,SIFT 描述子通过主成分分析和简单的阈值化操作变换为二值码.在文献[71]中,Affine-SIFT 通过调整 2 个照相机坐标方向参数,即纬度和经度,模拟原始图像的一组视角,有效覆盖了仿射变换的 6 个参数,实现了全仿射不变性.

从具有弱内部结构的区域抽取的 SIFT 特征具有很差的区分性,并可能使得图像检索性能下降.为了识别和移除这些特征,Dong 等[72]将 SIFT 描述子看作是一个取值范围 0~255 的离散随机变量的 128 个采样,然后采用熵作为测量标准滤除熵低的 SIFT 特征.

与像 SIFT 这样的浮点型特征不同,二值化特征被广泛探索,其可从感兴趣区域中直接抽取出来.近年来,二值化特征 BRIEF[73]和它的变体相继被提出,例如 ORB[74]、FREAK[75]和 BRISK[76],并在视觉匹配应用中吸引了极大的关注.这些二值化特征通过一些简单的强度差分测试得到,因此计算效率非常高.由于汉明距离的计算优势,基于 FAST 检测子[64]的二值化特征在大规模图像检索上具有很大潜力.Zhang 等[77]利用 DoG 检测子检测到局部区域中提取了一种新颖的极短二值化描述子.这种极短二值化描述子实现了快速匹配和索引.除此之外,遵循二值化 SIFT 方法[36],它避免了 BoW 模型中昂贵的码本训练和特征量化过程.较为全面的二值化描述子评估可参见文献[78].

除了如 SIFT 特征等局部区域中的梯度信息,边缘和颜色也能被用来表达成紧致的描述子,生成 Edge-SIFT[79]和 color-SIFT[80].作为一种二值化局部特征,Edge-SIFT[79]使用 Canny 边缘检测结果来描述一种局部区域.Zheng 等[68]从局部区域中抽取颜色名称特征,然后进一步变换到二值化特征以增强局部 SIFT 特征的区分性.

4.1.2 基于学习的特征

除了上面介绍的手工设计的视觉特征,我们还可以将数据驱动的方式学习特征用于图像检索.属性特征,原来在物体分类中使用,也可用来描述图像检索中的语义特征[81-83].属性单词表可通过人工的[84-85]或本体论式的[86]方式定义.对于每个属性,可使用核函数的分类器在有标签的训练图像集的多种低级视觉特征上进行训练,然后被用于预测不可见图像的属性评分[85-88].在文献[89]中,属性特征被用作一种语义一致的表达以辅助局部 SIFT 特征做图像检索.Karayev 等[90]学习分类器以预测图像类型,并将其应用到检索中,按照类型排列图像集合.属性特征的优点在于它提供了一种优雅的方式近似视觉语义,从而降低了语义鸿沟.但属性特征有 2 个不足.首先,无论采用手动或自动的方式,都很难定义一个属性单词表的完整集合,因此,基于有限属性单词表的表达在一个大的语义变化范围的图像数据集中是有偏差的.其次,由于需要在几千个属性类别上做分类,抽取语义特征计算代价高[81,86].主题模型,例如概率隐语义分析模型[91]和隐藏 Dirichlet 分布模型[92],也可用于学习语义特征表达做图像检索[93-94].

伴随着深度神经网络的突破性进展[65,95-96],近年来已经在很多领域见证了基于学习的特征的成功.使用深度架构,人们已经学习出接近人类识别过程的高层抽象[97].因此,人们采用深度神经网络从网络的不同层抽取特征.文献[98]从深度受限 Boltzmann 机的局部块中抽取出特征.作为深度神经网络的典型结构,深度卷积神经网络[99]已经在很多图像识别和检索任务[100]中显示出最优性能.在文献[101]中,针对不同的应用,包括基于内容的图像检索,作者基于深度卷积神经网络做了大量的实证分析.Razavian 等[102]深入研究 Alex-Net[99]和 VGG-Net[95],并探索了使用最后的卷积层 max pooling 响应作为图像表达进行图像检索.在文献[103]中,Alex-Net[99]的第 6 层激励被取出作为每幅图像的深度特征,并被融入传统的视觉特征,包括基于 SIFT 的 BoW 特征、HSV 直方图和 GIST 特征,用以计算图像相似性评分.

除了作为图像的全局表达,也能够以一种类似于局部特征的方式获得基于学习的特征[104].首先,采用无监督物体检测算法生成感兴趣的局部区域,例如选择性搜索[105]、objectness[106]和二范数梯度[107].这些算法生成大量的物体候选边界框.然后,在每个物体候选区域,抽取基于学习的特征.在文献[108]中,Sun 等采用 CNN 模型从通用物体检测子检测到的局部图像区域中抽取特征[107],然后将其应

用到图像检索中,获得了极好的性能.考虑到物体检测对于旋转变换的敏感性,Xie 等[104]提出旋转测试图像至 4 个不同角度,然后构建物体检测.具有最高物体检测评分的物体候选被用来抽取深度 CNN 特征[99].Tolias 等面向几何已知的重排序过程,生成卷积的局部最大响应特征(R-MAC)向量[109],扩展了积分图用于加速 max-pooling 操作.在文献[110]中,通过基于区域候选网络[111]的感兴趣区域选择器选择区域,R-MAC 描述子被扩展进行图像检索.

上面的方法均从分类任务中的深度学习模型中抽取基于学习的特征.因此,学习的特征可能不能很好地反映检索图像的视觉内容特性,这可能会限制检索的性能.因此,直接为检索任务训练深度学习模型是更受欢迎的,然而,这却很难实现,因为检索中的潜在图像类别很难定义或枚举.为了部分解决这个难点,Babenko 等[112]关注地标建筑物检索,使用与地标建筑物相关的类别调整在 ImageNet 上预训练的 CNN 模型.之后,在具有相似的视觉统计特性的检索数据集上,例如 Oxford Building 数据集[11],这种方法获得了有潜力的性能提升.为了摆脱对样本或类别标签的依赖,Paulin 等[113]提出一种基于卷积神经网络的无监督生成块级别的特征表达.文献[114]采用二值码的形式利用训练图像的相似性矩阵分解获得监督信息,最终的深度 CNN 模型能够以一种端到端的方式生成二值码.进一步地,Lai 等[115]提出利用深度神经网络把图像哈希为短的二值码,它的最优化是基于一种三元组排序损失.基于获得的短二值码作为图像表达可实现高效检索并降低存储复杂度.

4.2 视觉码本学习

通常,单幅图像中可抽取出成百上千的局部特征.为了实现紧致表达,高维的局部特征被量化到一个预先训练好的码本中的视觉单词,基于量化结果,一幅图像的局部视觉特征通过词袋(Bag-of-Visual-Words,BoW)模型[9]、VLAD[116]或 Fisher Vector[117],变换为一个定长向量.为了提前生成一个视觉码本,最直接的方式是通过 k-means 方法[9,12]对训练样本进行聚类,将聚类中心看作是视觉单词.由于局部特征维度很高并且训练样本集很大,训练百万甚至更大规模的大视觉码本需要极高的计算复杂度.为了解决这个问题,一种替换方法是采用层级 k-means[10],从线性律到对数律降低大尺寸的视觉码本生成的计算复杂度.

在标准 k-means 中,计算量最大的阶段是将每个特征分配到最近的聚类中心,这一步需要线性地比较所有的聚类中心.用最近邻搜索替换线性搜索可加速这个过程.基于这种观察,Philbin 等[11]提出一种近似 k-means 算法,使用随机 KD 树进行快速分配.Li 等[118]代替使用 k-means 生成视觉单词,预先定义一个半径随机采样种子点生成超球面,然后将种子点的超球面对应于视觉单词.在文献[119]中,Chu 等提出基于图密度建立视觉单词表,它采用图密度测量单词内相似性,并通过一个标量最大化估计方法生成视觉单词.

在 BoW 模型中,视觉码本作为一种媒介识别视觉单词 ID,这个 ID 可以被看作是量化或哈希的结果.换句话说,它使直接变换视觉特征到一个视觉单词 ID 而不显式定义视觉单词成为可能.基于这个观点,一些图像检索方法不需要直接训练就能生成一个虚拟的视觉码本.这些方法将一个局部特征变换到二值化特征,其中视觉单词 ID 被启发式定义.在文献[21]中,Zhang 等提出一个新的查询敏感的排序算法,用以排序基于 PCA 的二值哈希码,而后搜索 ϵ 邻域做图像检索.二值化特征使用局部敏感哈希策略生成,高比特位被用作视觉单词 ID,将相同 ID 的特征点分为一组.Zhou 等提出将一个 SIFT 描述子二值化为一个 256 bit 的二值化特征[36].无需训练码本,这个方法从 256 bit 的向量中选择 32 bit 作为码本,建立索引并检索.缺点是每个特征剩余的 224 bit 必须被存储在倒排索引表中,造成较大的内存消耗.相似地,Dong 等[72]提出使用 sketch embedding 方法[120]将一个 SIFT 描述子变换为 128 bit 的向量.然后,128 bit 向量被划分为 4 个不重叠的块,每个块被认为是一个键或后续索引的视觉单词.在文献[121]中,Zhou 等提出一个基于层级哈希的无码本训练的框架.为了确保特征匹配的召回率,大规模的层级哈希方法以一种层级的方式在局部特征的主成分上构建标量量化.

4.3 空间上下文嵌入

作为结构化的视觉内容的表达,视觉特征在图像平面上方向、尺度、关键点距离等空间上下文是相关的.引入上下文信息,视觉码本的区分能力能够被极大增强[26].与信息检索中的文本短语类似,在视觉单词上生成视觉短语是可行的.在文献[27,122]中,相邻的局部特征是相关的,可以用来生成高阶的视觉短语.视觉短语在内容表达上更加具有描述力.

很多算法在局部视觉特征中建立局部空间上下文.一些空间最近邻的弱空间一致性能够被用来滤除错误的视觉单词匹配.文献[9]通过校验具有15个最近邻定义的搜索区域的匹配特征收集正确的匹配.尽管这种弱约束有效,却对背景混乱的图像中的噪声很敏感.Zhang 等使用组距离度量,为组中的局部特征的空间上下文信息建模,生成语境视觉码本[28].Wang 等提出分别在描述子域和空间域中进行描述子上下文加权和局部特征空间上下文进行加权,以提升基于词汇树的方法的性能[123].描述子上下文加权通过统计描述子在词汇树中的量化路径的出现频率,降低信息量更少的描述子的权重,而空间上下文加权探索一些有效的空间上下文统计特性从而保留具有丰富描述力的局部特征.在文献[124]中,Liu 等通过在局部特征中嵌入空间上下文信息,建立了一种空间相关的单词表用于图像检索.

进一步地,多模态属性,即在一个相同的关键点上提取多种不同特征,被用于上下文哈希[125].在文献[126]中,几何最小哈希使用稀疏的局部几何信息构建可重复的哈希键以获得更加具有区分性的描述.在文献[17]中,Wu 等提出在 MSER 区域[61]捆绑局部特征.MSER 区域由区域及其边界中的强度函数的极值属性定义,通过基于分水岭的图像分割定义阈值范围可以检测出稳定 MSER 区域.捆绑特征通过共享视觉单词数目和匹配的视觉单词的相对排序进行比较.文献[63]在局部特征点的邻域提取顺序度量特征[127],然后构建局部空间一致性校验用于确定对应特征的顺序度量是否低于一个预先定义的阈值.

Cao 等提出了一种空间金字塔匹配方法[128]的推广策略,通过对2组有序的视觉单词进行线性投影和圆投影,并加以校准、均衡和分解等简单的直方图操作,对全局空间上下文信息建模,使特征具有平移、旋转和尺度不变性[129].

在人脸检索的场景中,上述码本生成方法可能不能抓取独特的面部特征.为了生成具有区分性的码本,Wu 等[130]提出使用一些具有不同姿势、表情和光照条件的训练人物样本生成基于个体的视觉单词表.一个视觉单词被定义为一个包含2种成分(分别是任务 ID 和位置 ID)的元组,并与多个样本相关.

4.4 特征量化

特征量化是在视觉码本定义之后,为每一个特征分配一个视觉单词的 ID.为了设计合适的分配函数,需要综合考虑量化精度、效率以及内存消耗.

最简单的方法是通过线性最近邻搜索,找出与特征最接近(最相似)的视觉单词,但是这种方法计算量较大.近似最近邻(ANN)搜索以牺牲精度为代价,提升了查找速度.文献[8]在 KD 树结构[131]中加入 best-bin-first 的策略,对查询图像的特征进行量化.文献[10]基于层级词汇树,从根节点开始逐层查找查询图像特征的最近邻.文献[132]提出了 KD 森林的近似算法,降低了时间复杂度. Muja 和 Lowe 使用 FLANN 库(www.cs.ubc.ca/research/flann),提出了优先查找 k-means 树算法用于可扩展的最近邻查找[133].文献[118]提出在随机播种码本上进行基于范围的查找来量化特征.尽管随机播种方法速度快,但是在训练数据上的偏差大,在大数据集上的检索精度有限[134].以上各种方法都采用硬量化,因此不可避免地引入了严重的量化误差.

码本将特征空间划分为一些不相交的区块,特征量化判定特征属于哪一个区块.当码本很大时,此时的特征空间划分是细粒度的,这意味着靠近区块边界的特征容易量化到不同的区块.当码本较小时,特征空间划分是粗粒度的,因为不相关的特征很可能被量化到相同的区块.这2种情况都会产生量化误差,并分别减低了特征匹配的召回率和准确率.因此必须折中考虑召回率和准确率以确定码本大小[10],或者引入某种限制以改善量化效果.

一些方法在采用大的码本的同时引入了软量化的方法以降低量化误差.一般而言,特征独立的软量化方法[15]将一个特征映射为多个视觉单词的加权组合.直观上,查询特征和数据库特征都可以进行软量化,但是,对数据库特征进行软量化会增加数倍存储开销.因此,软量化通常只在查询端进行[35].文献[35]基于 k-means 聚类得到的码本,以自底向上的方式,再进行 k-means 聚类,生成了2层的视觉词汇树,之后通过量化一个大的特征集合构建2层树节点之间的连接.该文提出的软量化基于距离比准则.

另一方面,其他方法采用相对较小的码本但增加了进一步的校验操作.文献[12]提出的汉明嵌入方法将 SIFT 特征映射到更低维的空间并训练一个中值向量,从而为每一个 SIFT 特征都生成一个二值特征码.每一个特征在量化之后都用该二值特征码进行匹配校验[54].文献[135]作为其变体,提出了非对称汉明嵌入方法以深入挖掘二值特征码的丰富信息.文献[136]也采用了类似的校验思想,利用单个

特征的中值生成了另一种不同的二值特征码.

上述量化方法都依赖于单个视觉码本.为了解决量化的块效应并提高召回率,多视觉码本应运而生[137-138].由于不同码本之间存在相关性,Zheng 等提出贝叶斯聚合方法以降低视觉码本交集特征的权重[139].他们从概率的角度为相关性问题建模并从图像和特征 2 个层面为视觉码本交集特征估计联合相似度.

局部特征的向量量化类似于近似最近邻搜索[58].已经有很多面向最近邻搜索问题的哈希算法被发表,例如:LSH[140-141]、多探针 LSH[142]、核化 LSH[56]、半监督哈希(SSH)[143]、谱哈希[57]、最小哈希[16]、迭代量化[144]、随机网格[145]、桶距离哈希(BDH)[146]、查询驱动的迭代近邻图搜索[147]以及线性距离保持哈希[148].然而大部分哈希算法都是应用于图像层次的全局特征如 GIST 和 BoW 特征,或仅用于局部特征层次的特征检索,很少有工作关注基于局部特征哈希的图像检索[22].主要原因是这些方法通常采用多个哈希表对每个特征进行索引,带来了巨大的内存消耗.LSH [141]、多探针 LSH[142]、核化 LSH[56]等方法需要将原始的数据库特征保存在内存中以计算它们与查询特征之间的距离,因此不适合大数据集的图像检索.另外,近似最近邻搜索致力于查找 k 个与查询最接近的数据,这违背了视觉特征匹配的基于范围的近邻搜索的本质.换句话说,给定一个查询特征,数据库中的目标特征个数是与查询特征相关的,且由查询特征的基于范围的近邻所确定.

文献[58]提出乘积量化产生指数级别的大码本而引入较小的内存消耗和近似最近邻搜索的时间消耗.乘积量化把特征空间分解为多个子空间的笛卡尔积,并对每个子空间进行独立量化.每个子空间的量化节都是一段短码,用这些短码建立查找表可以快速估计 2 个特征之间的欧式距离.然而由于乘积量化采用了穷举搜索,仍然不适用于大数据集上的图像检索[58].作为这个瓶颈问题的部分解,可以先采用 k-means 量化缩小搜索范围,再运用乘积量化[58].文献[149]在特征空间分解和码本建立 2 方面对乘积量化进行了优化,并提出了参数化和非参数化的 2 种方案.Zhou 等将特征匹配表示为 ϵ-近邻问题并用双重量化方法对其近似以进行快速索引和查询[134].他们对数据的每一个维度分别进行粗粒度和细粒度的量化并将各维度的量化结果串联起来.粗粒度量化结果用于构建索引,而细粒度量化结果用于生成二值特征码以进行匹配校验.文献[150]将高维的 SIFT 特征空间划分为规则网格.尽管在图像分类上有很好的效果,但是文献[15]证明了规则网格量化在大数据集的图像检索问题上比文献[10,15]的方法要糟糕得多.

4.5 特征聚合

当一张图像被表示为一个局部特征的集合,必须将这些局部特征聚合成一个固定长度的向量以进行图像之间相似度的计算.一般地,有 3 种方法可以实现局部特征聚合.

第 1 种,BoW 表达.每个特征被量化到最接近的视觉单词,其量化结果可以表示为一个高维的二值向量,非零值对应其量化到的视觉单词.将图像中所有特征量化的结果合并即得到 BoW 表达,该向量的维度是视觉码本的大小.由于视觉码本一般较大,因此图像的表达矩阵很稀疏,这使得倒排索引能够发挥很大用处.

第 2 种方法是 VLAD (Vector of Locally Aggregated Descriptor)[116],累加视觉单词与量化到该视觉单词的特征之间的残差,并将所有视觉单词对应的残差和串接起来,即可得到一个图像表征向量.VLAD 是一种紧凑的特征表达,并且继承了 SIFT 特征的特性包括平移不变性、旋转不变性和尺度不变性.文献[151]通过内归一化和多尺度 VLAD 表达提升了其性能;文献[152]对 VLAD 进行了深度分析;文献[153]结合三角嵌入和民主聚合策略拓展了 VLAD.更深入地,Tolias 等围绕 VLAD 提出了多种匹配方法[30].为了降低民主聚合的计算复杂度,Gao 等提出一种更快速的策略同时保持了相当的检索精度[154].文献[155]首先对局部特征进行稀疏编码,再通过最大池化聚合编码结果.Liu 等提出了构建 VLAD 的层级方法[156],通过引入隐藏层视觉词袋,残差向量的分布变得更加均匀,图像的特征表达更有区分力.

尽管对局部特征进行全局聚合得到了紧凑而有效的特征表达,然而 VLAD 特征对于解决图像部分遮挡和背景杂乱问题却没有很好的灵活性.为此,Liu 等[157]直接在图像层面上将关键点分组,再借助 VLAD[116]方法对每个组的局部特征聚合,从而得到可观的检索精度.

第 3 种是 Fisher Vector 表达[117,158-159].作为一个生成模型,给定图像的特征集合,Fisher Vector 用对

数似然函数的梯度来表示该图像[160].文献[117,161]采用高斯混合模型(GMM)聚合归一化的梯度向量.事实上,Fisher Vector 可以看成 BoW 和 VLAD 的衍生版本.一方面,如果只将关于混合高斯模型的权值的对数似然函数的梯度作为图像特征,那么 Fisher Vector 退化为软量化版本的 BoW.另一方面,如果只保留混合高斯模型的均值向量的对数似然函数的梯度,就得到了 VLAD 表达[58].Fisher Vector 和 VLAD 方法采用的混合高斯的数量或视觉码本都较小,得到的图像表达并不稀疏,以致于不适合使用倒排索引.因此,常对图像的表达向量进行降维和乘积量化[58]便于高效计算.

上述聚合手段基于局部手工特征,比如 SIFT.直观上,可以直接将这些方法移植到局部深度特征上.Gong 等[162]在多尺度下提取图像块的 CNN 特征,用 VLAD 方法对每一尺度下的特征进行聚合[37].文献[163]将最后一个卷积层的输出作为局部特征,他们证明单个局部特征已经具备较强的区分力,而聚合所有的局部特征将得到最好的性能.

5 数据库索引

索引是一种能迅速查询到目标图像的结构.由于检索的时间是非常重要的指标,随着数据库图像的不断增长,索引的重要性不言而喻.基于内容的图像检索通常采用 2 种索引方法:倒排索引和基于哈希的索引.接下来将分别介绍这 2 种索引方法.

5.1 倒排索引

受文本检索的启发,倒排索引[164]在大数据集图像检索领域也得到成功应用[9-12,14,17-18,165].本质上,倒排索引是一个稀疏矩阵的紧凑表达,其行和列分别表示图像和视觉单词.查询阶段,与查询图像包含共同视觉单词的数据库图像才会参与计算相似度,极大地提高了时间效率.

倒排索引中,每一个视觉单词都指向一个链表,链表中每一个单元都包含了图像 ID 等信息,甚至汉明码[12]、尺度、位置、方位等空间信息[11-13,18]也涵盖其中.文献[17]还记录特征在水平和垂直方向上的顺序,文献[123]的倒排索引包含了特征密度、平均对数尺度、平均方位差等空间统计量,文献[166]采用多 IDF 方法适应多种特征之间的相关性并将特征对应的二值特征码也存入倒排索引中.

倒排索引产生了很多变体.文献[42]面向基于轮廓的检索,分别在位置通道和方位通道量化图像边缘像素以建立倒排表,文献[68]提出多特征下的多索引结构,文献[70]在原始 SIFT 特征空间和二值 SIFT 空间交叉索引.

还有一些方法在倒排表中嵌入了语义信息.文献[167]通过图模型或矩阵分解把图像的表达分解为 2 部分,一部分用于降维,另一部分用于残差信息保持,而图像之间的相似度由这 2 部分决定.文献[89]通过语义属性删除了基于 SIFT 特征的倒排表中的不相关图像,同时插入语义相关图像,极大地增强了索引中特征的区分力.

为了提高召回率,数据库图像可能采用多个量化器以进行多次索引,比如 KD 树[66,168].文献[137]采用协同索引结构同时优化多个量化器.为了加速检索,文献[169]提出了 Q 索引,基于预定义的特征得分,剔除查询图像中不重要的特征同时只检索倒排表中较为重要的特征.针对并行检索,文献[170]在多个服务器上建立分布式索引,并将索引分布问题定义成一个学习问题以减少服务器之间的搜索延迟.

5.2 基于哈希的索引

当图像的特征表达向量不是稀疏的,如 GIST 特征和 VLAD 特征,倒排索引不再适用,而基于哈希的索引[171-175]得到广泛应用.最具代表性的是局部感知哈希(LSH)[176],使用多个随机映射哈希函数划分特征空间,当 2 个特征较为相似,它们发生冲突的概率较大.给定查询图像,基于哈希冲突可以筛选出一个候选列表,再通过精确的距离计算进行重排序.在文献[56]中,LSH 可以结合任何一种核函数,其时间复杂度可以是亚线性的.然而哈希方法的缺点是需要将数据库图像原始的特征保存在内存中.文献[177]将结合了图像外形和几何特性的特征图进行哈希索引,其空间复杂度是数据库图像特征个数的平方量级.文献[178]提出特征选择模型代替哈希方法以降低内存消耗.

倒排索引的内存消耗与图像特征向量中的非零元素个数成正比.为了进一步减少内存消耗,文献[179]将原始的 BoW 特征映射为多个最小 BOF 特征.这些最小 BOF 特征被进一步量化和索引.类似地,文献[16,180]用多个最小哈希函数把 BoW 特征映射到低维空间,每张图像需要保存在内存中的数据的比例是固定的.然而,尽管最小哈希[16,180]及其变体[126]能取得较高的检索精度,其召回率却不高,如果增加哈希表的个数,又会带来更多的内存消耗.

6 图像相关度评分

图像检索中,需要为每一个数据库中的图像分配一个得分并排序返回给用户.这种相似度得分一般定义成图像聚合特征之间的距离或者特征匹配时的投票得分.

6.1 基于距离的评分

将图像表示为1个定长向量之后,图像的相关性可以由2个向量之间的L_p归一化距离衡定:

$$D(\boldsymbol{I}_q,\boldsymbol{I}_m)=\left(\sum_{i=1}^{N}|q_i-m_i|^p\right)^{1/p},\quad(4)$$

$\boldsymbol{I}_q$和$\boldsymbol{I}_m$分别表示查询图像和数据库图像的N维特征表达向量.文献[10]证明了在BoW模型中,L_1归一化优于L_2归一化.文献[181]延伸了上述距离计算以测定图像的局部相似度并给出了优化方案.

在BoW模型中,为了区分不同视觉单词的重要性,词项频率(TF)和倒文档频率(IDF)被广泛采用[9,10,12,15,17].词项频率和倒文档频率加权之后,一般再进行L_p归一化.由于倒排索引的使用,距离的计算变得非常高效[10].

$$D(\boldsymbol{I}_q,\boldsymbol{I}_m)=\sum_{i=1}^{N}|q_i-m_i|^p=2+\sum_{i|q_i\neq0,m_i\neq0}(|q_i-m_i|^p-q_i^p-m_i^p).\quad(5)$$

然而L_p距离并不是最优的.文献[182]揭示了近邻关系不可逆问题,即一张图像不一定是它的近邻图像的近邻.为了解决这个问题,作者提出了上下文非相似性测度迭代修改图像之间的距离.文献[183]提出用概率模型来计算特征之间的相似度并引出查询自适应的计算方法.文献[184]直接通过扩散处理挖掘数据库图像的分布流形从而学得相似性测度.

文献[138]研究了BoW模型中共生及共消现象.共消,即1个视觉单词在2个BoW向量中对应的值都为0,这个问题可以通过减均值加以解决[138];视觉单词的共生会导致图像特征模式的重复计算,白化可以减弱其影响[138].这些操作还可以应用于VLAD模型中.

6.2 基于投票的评分

在基于局部特征的图像检索中,图像之间的相似度由特征的匹配程度决定,因此可以累加匹配特征的投票得到相似度得分,这种得分可直接排序而不需要归一化.

文献[13]简单地将特征匹配对数作为相似度得分;文献[35]将得分函数定义为查询和数据库图像共享视觉单词的TF-IDF的平方和,实际上即为BOF向量的内积;文献[17]将相似度定义为TF-IDF得分之和;文献[20]通过匹配特征集在文献[17]基础上进一步加权,加权项分为隶属度项和几何项,前者表示2组特征集共享的视觉词汇数量,后者则惩罚特征匹配的几何不一致性;文献[185-186]提出一种新颖的L_p范数IDF以拓展现有的IDF.

图像上下文空间信息对于图像匹配非常重要.文献[123]介绍了上下文加权机制结合IDF以提升视觉词汇树方法的性能,提出了描述子上下文加权(DCW)和空间上下文加权.文献[187]提出了基于某种变换的空间限制投票得分计算,变换空间被离散化并基于匹配特征的相对位置生成投票图,从而确定最优的变换.

文献[179]中每个特征被赋予一个二值特征码,图像距离被定义成所有匹配特征的二值特征码的汉明距离之和.为了保证不同变换下多个视觉对象的一致性,局部相似度由几何一致性匹配[188-189]的直方图的峰值决定.

在图像的视觉单词表达中,存在视觉单词"突发"现象,某些视觉单词在图像中出现的次数远高于统计均值,这不利于相似度得分的计算.对此,文献[190-191]提出了删除一对多匹配、加权弱化图像内/外"突发"等方法.

7 搜索重排序

初始查询结果可以通过发掘图像上下文信息[192-193]或增强初始查询等步骤得到改善.几何空间校验[11,13,18,126,194]、查询扩展[14,195]及查询融合[24]是提升查询精度最有效的3种后处理方案.接下来将详细介绍这3种方案.

7.1 几何空间校验

在基于局部特征的图像检索中,查询图像和数据库图像之间的特征对应由特征的相近性确立.典型地,如果2个局部特征量化到同一个视觉单词,则建立试探性对应.由于特征本身的歧义性和量化误差,错误的对应也常出现.基于此,特征的集合空间信息诸如空间位置、方位、尺度及特征的共生性等常被用来剔除错误的对应.在对应集合中,通常存在一个变换模型.仿射变换模型可以用来估计缩放、旋转、平移和视角变化等单应性变换.复杂情况可能存在多个单应性.

一些方法通过检验局部对应而直接预测变换模型,这些方法或基于类 RANSAC 算法[8,11,63,196],或基于霍夫投票方法[8,197].RANSAC 算法[198]的核心观点是产生对应集合的假设并鉴别出内点最多的模型.理论上,通过足够多的对应抽样和模型验证可以最大程度地恢复出变换模型,但是其计算量非常大.文献[11]引入了局部特征的区域外形,从而单个对应就可以产生一个假设,极大地减少了计算量.RANSAC 算法有 2 个缺点:第 1,RANSAC 算法需要参数以进行假设检验;第 2,RANSAC 算法的计算复杂度关于匹配的个数是平方量级的.

霍夫投票策略是在变换空间进行的[8,199],其计算复杂度与对应的个数成正比.文献[12]的霍夫投票是在尺度空间和方位空间进行的,基于 SIFT 特征下的对应,分别建立方位差和尺度差直方图,远离直方图峰值点的对应被认为是错误的对应.文献[20]基于特征对应之间的相对位移建立了二维霍夫投票空间,从而生成几何保持的视觉词组(GVP).如果不考虑霍夫直方图的内存代价,这种方法可以用来解决对尺度和方位的变化不变性.霍夫投票算法的缺点是对变换空间的划分粒度的定义不太灵活.为了解决这个问题,文献[197]受金字塔匹配模式[200]的启发,提出了霍夫金字塔匹配策略,并且这个策略的计算复杂度与对应的个数成线性关系.文献[199]在霍夫金字塔匹配的基础上对查询特征进行软量化,文献[194]提出成对几何匹配方法隐式地进行了空间校验,极大地降低了计算开销.

另一些方法则没有显式地处理变换模型.文献[9]利用局部特征组中的空间一致性来校验特征对应,文献[18]提出了对匹配特征对在水平和垂直 2 个方向上的相对坐标进行空间编码,并用该编码迭代地去除不满足空间一致性的匹配.文献[13,201]加入 SIFT 特征的方位和尺度对空间编码进行了延伸,提出了空间方格编码和空间扇形编码,能有效地解决图像平移、缩放、旋转等变换.文献[202]提出方向位置综合(COP)一致性图模型来度量 SIFT 特征对的相对空间一致性,通过检测特征匹配对集合最大的平均 COP,达到删除空间不一致的噪声特征的目的.

7.2 查询扩展

查询扩展亦借鉴自文本检索,用初始查询中排名靠前的结果生成新的查询.某些相关特征并未出现在初始查询图像中而出现在查询结果中,因此查询扩展可以使查询的特征表达更为丰富,从而提高了召回率.文献[14,195]讨论了平均查询扩展、传递闭包扩展、递归平均查询扩展、内扩展和外扩展等策略.

文献[23]将经过空间验证的图像作为正例,得分较低的图像作为反例,在线训练一个分类器,根据图像到分类决策面的距离进行初始结果重排.文献[203]在离线阶段建立了一个稀疏图结构连接潜在的相关图像,查询阶段采用 HITS 算法[204]进行关联性传播得到图像的排序结果.文献[205]更进一步地建立异构图模型并提出 2 种基于图结构的重排序方法,分别提高了召回率和准确率.文献[206]提出了空间查询扩展用于发掘普遍的视觉模式,这种查询扩展同时在视觉单词和图像 2 个层面进行.

作为查询扩展的一个特例,相关反馈[1]从 2000 年以来就得到持续关注[207-212],其应用也取得了极大成效.相关反馈依赖用户的标注区分相关和不相关图像并学得一个相似性测度. SVM[207-208] 和 boosting[213]是常用的学习算法.考虑到用户通常情况下不情愿去标注相关或不相关信息,用户的点击记录便成为极具价值的信息[31,214].文献[215-216]对相关反馈有更详尽的介绍.

7.3 检索融合

图像检索可以采用不同的图像特征和不同的算法[219].如果将不同的方法融合起来,优势互补,势必能得到更好的检索结果.很多融合方法都聚焦于排序阶段.文献[217]提出一种排序聚合算法以综合不同检索方法下得到的排序列表.文献[24]对每种检索方法的结果都建立无向图结构,再把所有的图结构融合成一张图,基于 PageRank 算法[218]或密度最大化策略得到最终的排序结果.

文献[103]在评分阶段进行检索融合.作者利用查询得分曲线下的面积来区分不同图像特征的表达效力,从而为每种特征分配一个权重,将不同特征下的得分加权相乘得到数据库图像与查询图像最终的相似度得分.

8 图像检索的评价指标

为了定量描述不同图像检索算法的精度与效率,必须收集标准数据集并定义衡量指标.此部分讨论图像检索研究中常用的有标注数据集以及干扰数据集,并描述图像目标检索中重要的衡量指标,如精确度、效率和内存占用等.

8.1 图像目标检索数据集

为了能够较好地体现出图像检索算法的可扩展性,标记数据集必须足够大.然而由于数据集收集过程中标注数据集是一个漫长的过程,因而现有的标记数据集都比较小,但是可以通过将其与达到百万规模的干扰数据集相结合来测试其可扩展性.现有的有标记数据集的目标都是特定的物体、场景以及部分重复的网络图片.一般来说,有标注的含有特定物体或场景的图片会经历各种变化,并且这些物体或者场景是在不同的光照强度、尺寸、角度、部分遮挡情况、压缩等条件下取得的.常用的标准数据集有UKBench 数据集[10]、Oxford 建筑物数据集[11]和 Holidays 数据集[12].MIR Flickr-1M 和 Flickr-1M 是 2 个不同的常作为干扰的数据集,各自均包含百万张图片.为了便于比较,表 1 中列举了图像目标检索中常用的数据集的相关信息.

UKBench 数据集(www.vis.uky.edu/~stewe/ukbench):该数据集包含 10 200 张图片,这些图片被分成 2 550 组.每组均包含四张不同视角或光照强度的描述同一物体的图片.所有的 10 200 张图片都作为检索目标,最后对它们的检索结果取平均.

Holidays 数据集(lear.inrialpes.fr/people/jegou/data.php):该数据集包含 1 491 张图片,这些图片被分成 500 组.每组图片均包含一个特定的物体或场景,并在不同的视角下拍摄.每组图片的第 1 张图片作为检索图片.

表 1 基于内容的图像检索中的常用数据集("mixed"表示该数据集包含了标注数据和干扰数据)

Table 1 General information of the popular retrieval datasets in CBIR(The "mixed" database type denotes that the corresponding dataset is a ground truth dataset mixed with distractor images)

图像库名称	图像库类型	数据库大小	查询图像数量	类别数量	图像分辨率(平均)
UKBench	Ground Truth	10 200	10 200	2 550	640×480
Holidays	Ground Truth	1 491	500	500	1 024×768
Oxford-5K	Mixed	6 053	55	11	1 024×768
Paris	Mixed	6 412	500	12	1 024×768
DupImage	Ground Truth	1 104	108	33	460×350
FlickrLogos-32	Mixed	8 240	500	32	1 024×768
INSTRE	Ground Truth	28 543	N/A	200	1 000×720
ZuBuD	Ground Truth	1 005	115	200	320×240
SMVS	Ground Truth	1 200	3 300	1 200	640×480
MIR Flickr-1M	Distractor	1 000 000	N/A	N/A	500×500
Flickr1M	Distractor	1 000 000	N/A	N/A	N/A

Oxford 建筑物数据集(www.robots.ox.ac.uk/~vgg/data/oxbuildings):该数据集由从 Flickr(www.flickr.com)网站上搜集到的 5 062 张牛津建筑物图片组成.这些图片已经被人工标注为 11 个不同的地标中的某一类,每一个地标都含有 5 个检索目标.因此共有 55 个检索目标.部分无关图片作为干扰项被加入到该数据集中.

Paris 数据集(www.robots.ox.ac.uk/~vgg/data/parisbuildings):该数据集由从 Flickr 数据集中选取的 6 412 张巴黎建筑物图片组成.该数据集共有 500 张检索图片.

DupImage 数据集(pan.baidu.com/s/1jGETFUm):该数据集包含 1 104 张图片,被分成 33 组.每一组的内容为一个图标或一个插画,比如肯德基图标、美国哥特式绘画、蒙娜丽莎等.从中选取 108 张图片作为检索图片.

FlickrLogos-32 数据集(www.multimedia-computing.de/flickrlogos):该数据集由从 Flickr 数据集中选取的 32 个商标图标组成.该数据集被分成训练部分、验证部分和测试部分.8 240 张图片中有 6 000 张图片不包含图标,将其作为干扰项.

INSTRE 数据集(vipl.ict.ac.cn/isia/instre):该数据集包含 INSTRE-S 和 INSTRE-m 2 个部分[221].前者包含 200 类 23 070 张图片,后者包含 5 473 张图片,每张图片包含 100 类目标中的 2 个实例.

ZuBuD 数据集(www.vision.ee.ethz.ch/showroom/zubud/index.en.html):该数据集包含苏黎世的 201 个建筑物共 1 005 张图片,每一建筑物有 5 个不同的视角[222].数据集中加入了 115 张无关图片,这些图片在数据集中找不到相关图片.图片的分

辨率为 320×240.

Stanford Mobile Visual Search 数据集(purl.stanford.edu/rb470rw0983):该数据集为手机相机拍摄的照片,比如 CD、书本、户外建筑物、名片、博物馆的艺术品、唱片等.数据集中共有 3 300 张检索图片.

MIR Flickr-1M 数据集(medialab.liacs.nl/mirflickr/mirflickr1m):该数据集为干扰数据集,由 Flickr 数据集中随机选取的 100 万张图片组成,每张图片经过缩放后不大于 500×500.

Flickr1M 数据集(bigimbaz.inrialpes.fr/herve/siftgeo1M)是另一个干扰数据集,包含了 Flickr 数据集中 100 万张图片的 SIFT 特征.该数据集中不包含原始图片.

8.2 图像目标检索评价指标

多媒体图像目标检索系统中,精确度、效率和内存占用是 3 个重要的评价指标.通常检索算法都希望在最小的牺牲其中 2 个指标的情况下提升另一个指标.

1) 精确度.为了定量地描述检索结果,根据相关程度对数据集图片进行分类,并依据数据库图片的返回顺序来计算精确度得分.不同的相关水平具有不同的距离.实际中只使用 2 种相关水平:相关与不相关.平均精确度用来衡量单张图片的检索结果.平均精确度结合了精确率和召回率.精确率表示检索得到的前 k 张图片中正确结果的比例.召回率表示检索结果中的正确结果与真实正确结果的比例.一般来说,如果一个检索系统的精确度降低,则其检索结果中的正确结果以及召回率会上升.如式(6)所示,当一张相关图像被检索到时,将其返回序号取平均作为平均精确度.为了描述多张检索图片的检索结果,将每一张检索图片的平均精确度取平均,得到平均精确度均值.

$$A_{\mathrm{P}}=\frac{\sum_{k=1}^{n}P(k)\times \mathrm{rel}(k)}{R}, \tag{6}$$

其中 R 代表当前检索图片的相关图片数量,$P(k)$ 代表前 k 个检索结果的平均精确度,$\mathrm{rel}(k)$ 代表第 k 个检索结果是否为相关图片,若为相关图片,则为 1,否则为 0,n 代表检索结果的总数量.

当使用多种相关度级别的时候,使用式(7)所示的归一化搜索引擎质量指标来衡量检索结果.

$$N_{\mathrm{DCG}}=\frac{1}{N}\left(r_1+\sum_{k=2}^{n}\frac{f(r_k)}{\log_2(k)}\right), \tag{7}$$

其中 n 代表检索得到的图片的数量,r_k 代表相关度级别,$f(*)$ 是调整不同相关性水平的函数,N 表示归一化项,以确保在检索结果理想时该指标的结果为 100%.$f(*)$ 的常用定义包括 $f(x)=x$ 和 $f(x)=2^x-1$,后者强调检索高度相关的图像.

除了以上评价指标,特定数据集还有特定的评价指标.在 UKBench 数据集中,由于每一个检索目标都有 4 张相关图片,因此使用 N-S 得分,即用前 4 个返回结果中正确结果个数的均值来反映检索精确度.

2) 计算效率.图像目标检索的计算效率包括建立码本花费的时间、视觉特征索引花费的时间以及检索花费的时间.前 2 项是离线进行的,最后一项是在线进行的,都希望花费时间尽可能短.在线的检索过程需要具有实时性.

3) 内存占用.多媒体图像目标检索系统中,内存占用指的是在线检索过程阶段占用的内存.一般来说,内存主要用于聚类器及需要在检索开始前提前导入内存的数据集索引文件.常用的聚类算法是基于树形结构的,比如分级单词树和随机森林等,这些包含几百万个视觉单词的树一般占用几百兆字节空间.倒排表占用的内存空间与数据集大小成正相关.当用局部特征表示数据集图片且局部特征已被索引到时,倒排表占用的空间与局部特征占用的内存空间成正相关.

9 未来的研究方向探讨

过去几十年来,产生了许多新方法以改善图像目标检索系统,然而仍然有很大改善空间.接下来,我们将讨论未来几十年的研究方向.

9.1 收集标记数据集

在多媒体和计算机视觉领域,往往是特定的任务驱使新的标记数据集产生.构建数据集初期,研究人员不断地提出经典的方法刷新检索精确度并解决研究问题.但在此过程中,数据集的过拟合可能会阻碍算法上的突破.随着对研究问题有了更好的理解并对其有了更加明确的定义,现有数据集的不足逐渐显现,因此需要采集新的数据集.新数据集的标记应足够准确,从而消除一些图像内容相关性上存在的二义性问题,比如商品图标数据集等.同时,数据集应足够大从而将其与图像分类问题区分开.

9.2 意图导向的查询生成与选择

意图鸿沟是基于内容的图像检索中首要也是最

大的一个挑战.一个简单的查询问题,例如,彩色图或草图,在大部分场合下仍然无法反映用户意图,使得检索结果不理想.除了传统的查询方式,用户指定具体检索意图可大大降低后期的检索难度.考虑到可能用户参与检索过程的意愿低,可以设计方便的查询界面接口以尽可能减少用户参与.例如,对用户而言,在用于检索的示例图像中指定感兴趣的区域,或指出预期的结果是部分重复的,或指示类似的空间颜色和纹理结构等,则是很容易的.也可以预测可能的意图并与用户确认.总而言之,相比于被动预测用户的意图,更佳的办法是使用户积极参与到检索过程中.

在图像检索中,检索效果会受到查询图像的影响.如何选择一个最适合检索的查询图像是一个非常重要的问题.查询图像的质量的相关因素包括分辨率、噪声、仿射变换、背景的杂乱程度等.在移动搜索的场合下,可以让用户拍摄更好的照片以获得更好的查询图片.在服务器端,可以设计检索质量自动评估方法[223-224]从初始检索结果中选取高精度结果作为潜在的候选结果.

9.3 面向检索的深度学习

尽管基于内容的视觉检索取得了较大进展,但语义感知检索与视觉内容仍存在巨大的鸿沟.因为目前用于图像表达的特征都是手动设计的,所以无法捕捉语义信息.由于多媒体视觉数据的多样性,现有的方法是无监督的.为了解决语义感知检索方面的难题,可使用可扩展监督或半监督学习进行语义学习,以提高基于内容的视觉检索的性能.大规模视觉识别的深度学习的成功[95-96,99,225]已经表明其具备这样的潜力.

将现有的深度学习方法运用于基于内容的图像检索,首先需要解决2个重要的问题.第一,深度学习获得的图像表达应灵活多变并且对各种常见的变换具有不变性,如旋转变换和缩放变换.由于现有的深度学习特征是将图像与各向异性卷积滤波器进行卷积获取的,所得特征图对大幅度的旋转变换和缩放变换不具有不变性.目前仍无法确定能否通过增加训练样本来解决.第二,由于基于内容的视觉目标检索中特别强调计算效率和内存占用,在设计深度学习网络时需考虑这些限制因素.例如,紧凑的二进制语义哈希编码[59,65]和稀疏的语义向量均可以用于表示图像,但因为后者在距离计算和内存占用2方面都有较高的效率,所以更适合于倒排索引结构.

9.4 无监督数据库融合

传统的基于内容的图像检索算法和系统中,数据库图像被独立处理,其潜在的相关性信息则没有被考虑.主要原因是通常没有数据库图像的标签信息而且潜在的类别数量是无限的.这些问题限制了复杂监督学习在图像目标检索算法中的应用.不过,只要数据库足够大,很可能存在一些图像子集,而每个子集中的图像可能与其他子集中的图像相关.因此,在离线阶段需要使用无监督技术发现这些子集的相关关系.如果将每个数据库图像作为节点,将图像之间的相关性程度当作连接节点的边权值,则可以用图结构表示所有的数据库图像.那么,子集相关问题可以视为子图发现问题.另一方面,在实践中,新图像可以增加至原图中.离线阶段的最终结果可使在线查询获得更好的检索结果.

9.5 跨模态检索

上述讨论中,我们专注于图像目标检索.但是,除了视觉特征,还有其他非常有用的信息,如网页中图像的文字信息、用户在使用搜索引擎时的搜索日志、视频中的语音信息等.这些多模态信息是互补的,有利于协同识别图像和视频的视觉内容.因此,可以使用不同的模型探索跨模态信息检索并整合这些信息.基于多模态信息表达、特征量化、建立索引、搜索重排序将成为新的研究课题.

9.6 端到端的检索框架

如上节所述,检索框架涉及多个模块,包括特征提取、码本学习、特征量化、图像索引等.这些模块均对每一个检索任务单独设计和独立优化.此外,若研究对象为深度学习中卷积神经网络的结构,我们可以在BoW模型和CNN模型中找到一个非常密切的类比.卷积滤波器在CNN模型中的使用方式与BoW模型中的码本视觉单词类似.图像块和卷积滤波器的卷积结果本质上是软量化结果,它们的极大值池化操作类似于BoW模型中的聚类操作.只要学习到的特征向量是稀疏矩阵,就可以有效地采用倒排索引结构建索引图像数据库.与BoW模型不同的是,CNN模型中上述模块是针对图像分类的任务优化.类似地,我们也可以采取端到端方案,将图像作为框架的输入,输出索引的特征,并使用传统的关键检索相关模块进行协同优化.

9.7 图像目标检索与社交媒体

与传统的无结构网络多媒体数据不同,近几年,

社交媒体平台上分享了大量的社交媒体数据.代表性的社交媒体平台,如 Facebook、Twitter、维基百科、LinkedIn、Pinterest 等.社交媒体上含有海量的多态信息,这些信息既体现社会文化背景和潮流趋势,也揭示个人的情感和行为特征等.基于内容的图像检索技术,在用户创建的内容中,视觉数据用途广泛,即可以发掘和理解潜在的社区关系,帮助了解个体用户的行为,提供产品推荐服务,还可以进行人群情绪监督和预警.

9.8 公开挑战赛

由于数据的部署结构和可获得性不同,学术界的研究和工业界的应用存在巨大的鸿沟.为了解决这个问题,应鼓励科研人员参与一些工业界的项目,并在实际场景中解决遇到的关键问题.过去的 5 年中,已经产生类似的项目,比如微软图像检索挑战赛和阿里巴巴大规模图像检索挑战赛.这些挑战赛不仅会促使学术界研究的发展,还可以解决现实中的各种问题,相信将来会有越来越多的挑战赛或类似项目.

10 全文总结

本文主要总结近年来图像目标检索的发展.首先,重点阐述检索目标的产生、图像表示、图像索引、检索得分的重排序等检索框架的关键模块,然后,分别讨论每一个模块的关键问题以及一些代表性研究阶段和方法,最后,扩展讨论了未来 8 个可能的提高检索性能的研究方向.

参考文献

References

[1] Rui Y,Huang T S,Ortega M,et al.Relevance feedback:A power tool for interactive content-based image retrieval[J].IEEE Transactions on Circuits and Systems for Video Technology,1998,8(5):644-655

[2] Alzubi A,Amira A,Ramzan N.Semantic content-based image retrieval:A comprehensive study[J].Journal of Visual Communication and Image Representation,2015,32:20-54

[3] Li X R,Uricchio T,Ballan L,et al.Socializing the semantic gap:A comparative survey on image tag assignment,refinement,and retrieval[J].ACM Computing Surveys,2016,49(1):14

[4] Lin Z J,Ding G G,Hu M Q,et al.Semantics-preserving hashing for cross-view retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition,2015:3864-3872

[5] Smeulders A W M,Worring M,Santini S,et al.Content-based image retrieval at the end of the early years[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(12):1349-1380

[6] Lew M S,Sebe N,Djeraba C,et al.Content-based multimedia information retrieval:State of the art and challenges[J].ACM Transactions on Multimedia Computing,Communications,and Applications,2006,2(1):1-19

[7] Liu Y,Zhang D S,Lu G J,et al.A survey of content based image retrieval with high-level semantics[J].Pattern Recognition,2007,40(1):262-282

[8] Lowe D G.Distinctive image features from scale invariant keypoints[J].International Journal of Computer Vision,2004,60(2):91-110

[9] Sivic J,Zisserman A.Video Google:A text retrieval approach to object matching in videos[C]// IEEE International Conference on Computer Vision and Pattern Recognition,2003:1470-1477

[10] Nister D,Stewenius H.Scalable recognition with a vocabulary tree[C]// IEEE Conference on Computer Vision and Pattern Recognition,2006:2161-2168

[11] Philbin J,Chum O,Isard M,et al.Objectret rieval with large vocabularies and fast spatial matching[C]// IEEE Conference on Computer Vision and Pattern Recognition,2007:1-8

[12] J'egou H,Douze M,Schmid C.Hamming embedding and weak geometric consistency for large scale image search[C]// European Conference on Computer Vision,2008:304-317

[13] Zhou W G,Li H Q,Lu Y J,et al.Large scale image search with geometric coding[C]// ACM International Conference on Multimedia,2011:1349-1352

[14] Chum O,Philbin J,Sivic J,et al.Total recall:Automatic query expansion with a generative feature model for object retrieval[C]// IEEE International Conference on Computer Vision,2007:1-8

[15] Philbin J,Chum O,Isard M,et al.Lost in quantization:Improving particular object retrieval in large scale image databases[C]// IEEE Conference on Computer Vision and Pattern Recognition,2008:1-8

[16] Chum O,Philbin J,Zisserman A.Near duplicate imagedetection:Min-hash and TF-IDF weighting[C]// British Machine Vision Conference,2008,3:4

[17] Wu Z,Ke Q F,Isard M,et al.Bundling features for large scale partial-duplicate web image search[C]// IEEE Conference on Computer Vision and Pattern Recognition,2009:25-32

[18] Zhou W G,Lu Y J,Li H Q,et al.Spatial coding for large scale partial-duplicate web image search[C]// ACM International Conference on Multimedia,2010:511-520

[19] Chum O,Mikulik A,Perdoch M,et al.Total recall Ⅱ:Query expansion revisited[C]// IEEE Conference on Computer Vision and Pattern Recognition,2011:889-896

[20] Zhang Y M,Jia Z Y,Chen T.Image retrieval with geometry-preserving visual phrases[C]// IEEE Conference on Computer Vision and Pattern Recognition,2011:809-816

[21] Zhang X,Zhang L,Shum H-Y.QsRank:Query-sensitive hash code ranking for efficient ϵ-neighbor search[C]//

IEEE Conference on Computer Vision and Pattern Recognition,2012:2058-2065
[22] He J F,Feng J Y,Liu X L,et al.Mobile product search with bag of hash bits and boundary reranking[C]//IEEE Conference on Computer Vision and Pattern Recognition,2012:3005-3012
[23] Arandjelovic R, Zisserman A. Three things everyone should know to improve object retrieval[C]//IEEE Conference on Computer Vision and Pattern Recognition,2012:911-2918
[24] Zhang S T,Yang M,Cour T,et al.Query specific fusion for image retrieval[C]//European Conference on Computer Vision,2012:660-673
[25] Tian Q,Zhang S L,Zhou W G,et al.Building descriptive and discriminative visual codebook for large-scale image applications[J]Multimedia Tools and Applications,2011,51(2):441-477
[26] Zhou W G,Li H Q,Lu Y J,et al.Large scale partial-duplicate image retrieval with bi-space quantization and geometric consistency[C]//IEEE International Conference Acoustics Speech and Signal Processing, 2010:2394-2397
[27] Zhang S L,Tian Q,Hua G,et al.Descriptive visual words and visual phrases for image applications[C]//ACM International Conference on Multimedia,2009:75-84
[28] Zhang S L,Huang Q M,Hua G,et al.Building contextual visual vocabulary for large-scale image applications[C]//ACM International Conference on Multimedia,2010:501-510
[29] Zhou W G,Tian Q,Lu Y J,et al.Latent visual context learning for web image applications[J].Pattern Recognition,2011,44(10/11):2263-2273
[30] Tolias G,Avrithis Y,J'egou H.To aggregate or not to aggregate:Selective match kernels for image search[C]//International Conference on Computer Vision, 2014:1401-1408
[31] Zhang L,Rui Y.Image search from thousands to billions in 20 years[J].ACM Transactions on Multimedia Computing, Communications, and Applications, 2013, 9(1):36
[32] Tang X O,Liu K,Cui J Y,et al.Intent search:Capturing user intention for one-click internet image search[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,34(7):1342-1353
[33] Moghaddam B,Tian Q,Lesh N,et al.Visualization and user-modeling for browsing personal photo libraries[J].International Journal of Computer Vision,2004,56(1/2):109-130
[34] Datta R,Joshi D,Li J,et al.Image retrieval:Ideas,influences,and trends of the new age[J].ACM Computing Surveys,2008,40(2):5
[35] J'egou H,Douze M,Schmid C.Improving bag-of-features for large scale image search[J].International Journal of Computer Vision,2010,87(3):316-336
[36] Zhou W G,Lu Y J,Li H Q,et al.Scalar quantization for large scale image search[C]//ACM International Conference on Multimedia,2012:169-178
[37] Cao Y,Wang H,Wang C H,et al.Mindf inder:Interactive sketch-based image search on millions of images[C]//ACM International Conference on Multimedia, 2010:1605-1608
[38] Xiao C C,Wang C H,Zhang L Q,et al.Sketch-based image retrieval via shape words[C]//ACM International Conference on Multimedia Retrieval,2015:571-574
[39] Sousa P,Fonseca M J.Sketch-based retrieval of drawings using spatial proximity[J]//Journal of Visual Languages & Computing,2010,21(2):69-80
[40] Fonseca M J,Ferreira A,Jorge J A.Sketch-based retrieval of complex drawings using hierarchical topology and geometry[J].Computer-Aided Design, 2009, 41(12):1067-1081
[41] Liang S,Sun Z X.Sketch retrieval and relevance feedback with biased SVM classification[J].Pattern Recognition Letters,2008,29(12):1733-1741
[42] Cao Y,Wang C H,Zhang L Q,et al.Edgel index for large scale sketch-based image search[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2011:761-768
[43] Wang J D,Hua X-S.Interactive image search by color map[J].ACM Transactions on Intelligent Systems and Technology,2011,3(1):1-23
[44] Xu H,Wang J D,Hua X-S,et al.Image search by concept map[C]//International ACM SIGIR Conference on Research and Development in Information Retrieval,2010:275-282
[45] Xu H,Wang J D,Hua X-S,et al.Interactive image search by 2D semantic map[C]//International Conference on World Wide Web,2010:1321-1324
[46] Lan T,Yang W L,Wang Y,et al.Image retrieval with structured object queries using latent ranking SVM[J].European Conference on Computer Vision,2012:129-142
[47] Kim G,Moon S,Sigal L.Ranking and retrieval of image sequences from multiple paragraph queries[C]//IEEE Conferenceon Computer Vision and Pattern Recognition,2015:1993-2001
[48] Wengert C,Douze M,J'egou H.Bag-of-colors for improved image search[C]//ACM International Conference on Multimedia,2011:1437-1440
[49] Xie J,Fang Y,Zhu F,et al.Deepshape:Deep learned shape descriptor for 3d shape matching and retrieval[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:1275-1283
[50] Wang F,Kang L,Li Y.Sketch-based 3D shape retrieval using convolutional neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:1875-1883
[51] Bai S,Bai X,Zhou Z C,et al.Gift:A real time and scalable 3d shape search engine[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2016:5023-5032
[52] Park M,Jin J S,Wilson L S.Fast content-based imageretrieval using quasi-Gabor filter and reduction of image feature dimension[C]//IEEE Southwest Symposium on Image Analysis andInterpretation,2002:178-182
[53] Wang X Y,Zhang B B,Yang H Y.Content-based image retrieval by integrating color and texture features[J].

Multimedia Tools and Applications, 2014, 68 (3): 545-569
[54] Wang B,Li Z W,Li M J,et al.Large-scale duplicate detection for web image search[C]// IEEE International Conference on Multimedia and Expo,2006:353-356
[55] Siagian C,Itti L.Rapid biologically-inspired scene classification using features shared with visual attention[J]. IEEETransactions on Pattern Analysis and Machine Intelligence,2007,29(2):300-312
[56] Kulis B,Grauman K.Kernelized locality-sensitive hashing for scalable image search[C]// IEEE International Conference on Computer Vision,2009:2130-2137
[57] Weiss Y,Torralba A,Fergus R.Spectral hashing[C]// International Conference on Neural Information Processing Systems,2008:1753-1760
[58] J'egou H, Douze M, Schmid C. Product quantization for nearest neighbor search [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2011,33(1): 117-128
[59] Torralba A,Fergus R,Weiss Y.Small codes and large image databases for recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition,2008:1-8
[60] Lowe D G. Object recognition from local scale-invariant features [C] // IEEE International Conference on Computer Vision,1999,2:1150-1157
[61] Matas J,Chum O,Urban M,et al.Robust wide baseline stereo from maximally stable extremal regions[J].Image- and Vision Computing,2004,22(10):761-767
[62] Mikolajczyk K,Schmid C.Scale & affine invariant interest point detectors[J].International Journal of Computer Vision,2004,60(1):63-86
[63] Xie H T,Gao K,Zhang Y D,et al.Efficient feature detection and effective post-verification for large scalenear-duplicate image search[J].IEEE Transactions on Multimedia,2011,13(6):1319-1332
[64] Rosten E, Porter R, Drummond T. Faster and better: A machine learning approach to corner detection[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2010,32(1):105-119
[65] Krizhevsky A,Hinton G E.Using very deep autoencoders for content-based image retrieval[C]// European Symposium on Artificial Neural Networks,2012
[66] Wu Z,Ke Q F,Sun J,et al.A multi-sample,multitree approach to bag-of-words image representation for image retrieval[C]// IEEE International Conference on Computer Vision,2009:1992-1999
[67] Bay H, Tuytelaars T, Van Gool L. SURF: Speeded up robust features[C]// European Conference on Computer Vision,2006:404-417
[68] Zheng L,Wang S J,Liu Z Q,et al.Packing and padding: Coupled multi-index for accurate image retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition,2014:1947-1954
[69] Zhou W G, Li H Q, Hong R C, et al. BSIFT: Towards data-independent codebook for large scale image search [J]. IEEE Transactions on Image Processing, 2015, 24 (3):967-979
[70] Liu Z,Li H Q,Zhang L Y,et al.Cross-indexing of binary SIFT codes for large-scale image search[J].IEEE Transactions on Image Processing,2014,23(5):2047-2057
[71] Yu G S,Morel J M.ASIFT:An algorithm for fully affine invariant comparison [J]. Image Processing on Line, 2011,1:2105-1232
[72] Dong W, Wang Z, Charikar M, et al. High-confidence near-duplicate image detection[C]// ACM International Conferenceon Multimedia Retrieval,2012:1
[73] Calonder M,Lepetit V,Strecha C,et al.BRIEF:Binary robust independent elementary features [C] // European Conference on Computer Vision,2010:778-792
[74] Rublee E,Rabaud V,Konolige K,et al.ORB:An efficient alternative to SIFT or SURF [C] // International Conference on Computer Vision,2011:2564-2571
[75] Alahi A, Ortiz R, Vandergheynst P. FREAK: Fast retina keypoint[C]// IEEE Conference on Computer Vision and Pattern Recognition,2012:510-517
[76] Leutenegger S, Chli M, Siegwart R Y. BRISK: Binary robust invariant scalable keypoints [C] // International Conference on Computer Vision,2011:2548-2555
[77] Zhang S L,Tian Q,Huang Q M,et al.USB:Ultrashort binary descriptor for fast visual matching and retrieval[J]. IEEE Transactions on Image Processing, 2014, 23(8): 3671-3683
[78] Madeo S, Bober M. Fast, compact and discriminative: Evaluation of binary descriptors for mobile applications [J]. IEEE Transactions on Multimedia, 2016, 19(2): 221-235
[79] Zhang S L,Tian Q,Lu K,et al.Edge-SIFT:Discriminative binary descriptor for scalable partial-duplicate mobile search[J]. IEEE Transactions on Image Processing, 2013,22(7):2889-2902
[80] Van De Sande K,Gevers T,Snoek C G.Evaluating color descriptors for object and scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2010,32(9):1582-1596
[81] Douze M,Ramisa A,Schmid C.Combining attributes and Fisher vectors for efficient image retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2011:745-752
[82] Zhao S C, Yao H X, Yang Y, et al. Affective image retrieval via multi-graph learning [C] // ACM International Conference on Multimedia,2014:1025-1028
[83] Tao R,Smeulders A W,Chang S F.Attributes and categories for generic instance search from one example[C]// IEEE Conference on Computer Vision and Pattern Recognition,2015:177-186
[84] Farhadi A,Endres I,Hoiem D,et al.Describing objects by their attributes[C]// IEEE Conference on Computer Vision and Pattern Recognition,2009:1778-1785
[85] Khan F S,Anwer R M,Van De Weijer J,et al.Color attributes for object detection [J]. IEEE Conference on Computer Vision and Pattern Recognition, 2012: 3306-3313
[86] Torresani L, Szummer M, Fitzgibbon A. Efficient object category recognition using classemes [C] // European Conference on Computer Vision,2010:776-789
[87] Jia D,Berg A C,Li F F.Hierarchical semantic indexing

for large scale image retrieval[C]//IEEE Conference on Computer Vision and Pattern Recognition,2011:785-792
[88] Cai J J,Zha Z J,Wang M,et al.An attribute assistedre ranking model for web image search [J]. IEEE Transactions on Image Processing,2015,24(1):261-272
[89] Zhang S L,Yang M,Wang X Y,et al.Semantic-awareco-indexing for image retrieval[J].IEEE International Conferenceon Computer Vision,2013,37(12):1673-1680
[90] Karayev S,Trentacoste M,Han H,et al.Recognizing image style[J].arXiv e-print,2013,arXiv:1311.3715
[91] Hofmann T.Unsupervised learning by probabilistic latent semantic analysis[J].Machine Learning,2001,42(1/2):177-196
[92] Blei D M,Ng A Y,Jordan M I.Latent dirichlet allocation [J].Journal of Machine Learning Research,2003,3:993-1022
[93] Hörster E,Lienhart R,Slaney M.Image retrieval on large scale image databases [C] // ACM International Conference on Image and Video Retrieval,2007:17-24
[94] Lienhart R,Slaney M.PLSA on large scale image databases[C]//IEEE International Conference on Acoustics,Speechand Signal Processing,2007:1217-1220
[95] Simonyan K,Zisserman A.Very deep convolutional networks for large-scale image recognition[J].arXiv e-print,2014,arXiv:1409.1556
[96] Szegedy C,Liu W,Jia Y Q,et al.Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition,2014:1-9
[97] Bengio Y.Learning deep architectures for AI[J].Foundations and trends in Machine Learning,2009,2(1):1-127
[98] Hörster E,Lienhart R.Deep networks for image retrieval on large-scale databases[C]//ACM International Conference on Multimedia,2008:643-646
[99] Krizhevsky A,Sutskever I,Hinton G E.ImageNet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems,2012:1097-1105
[100] Razavian A S,Azizpour H,Sullivan J,et al.CNN features off-the-shelf:An astounding baseline for recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition,2014:512-519
[101] Wan J,Wang D Y,Hoi S C H,et al.Deep learning for content-based image retrieval: A comprehensive study [C]//ACM International Conference on Multimedia,2014:157-166
[102] Razavian A S,Sullivan J,Carlsson S,et al.Visual instance retrieval with deep convolutional networks[J].arXiv e-print,2014,arXiv:1412.6574
[103] Zheng L,Wang S J,Tian L,et al.Query-adaptive late fusion for image search and person reidentification[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:1741-1750
[104] Xie L X,Hong R C,Zhang B,et al.Image classification and retrieval are ONE [C] // ACM International Conference on Multimedia Retrieval,2015:3-10
[105] Uijlings J R R,Van De Sande K E,Gevers T,et al.Selective search for object recognition [J]. International Journal of Computer Vision,2013,104(2):154-171
[106] Alexe B,Deselaers T,Ferrari V.Measuring the objectness of image windows[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,34(11):2189-2202
[107] Cheng M M,Zhang Z,Lin W Y,et al.Bing:Binarized normed gradients for objectness estimation at 300 fps[C]//IEEE Conference on Computer Vision and Pattern Recognition,2014:3286-3293
[108] Sun S Y,Zhou W G,Tian Q,et al.Scalable object retrieval with compact image representation from generic object regions [J]. ACM Transactions on Multimedia Computing,Communications,and Applications,2015,12(2):29
[109] Tolias G,Sicre R,J'egou H.Particular object retrieval with integral max-pooling of CNN activations[J].arXiv e-print,2015,arXiv:1511.05879
[110] Gordo A,Almazan J,Revaud J,et al.Deep image retrieval:Learning global representations for image search [C]//European Conference on Computer Vision,2016:241-257
[111] Ren S Q,He K M,Girshick R,et al.Faster R-CNN:Towards real-time object detection with region proposal networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149
[112] Babenko A,Slesarev A,Chigorin A,et al.Neural codes for image retrieval[C]//European Conference on Computer Vision,2014:584-599
[113] Paulin M,Douze M,Harchaoui Z,et al.Local convolutional features with unsupervised training for image retrieval[C]//IEEE International Conference on Computer Vision,2015:91-99
[114] Xia R K,Pan Y,Lai H J,et al.Supervised hashing for image retrieval via image representation learning[C]//AAAI Conference on Artificial Intelligence,2014:2156-2162
[115] Lai H J,Pan Y,Liu Y,et al.Simultaneous feature learning and hash coding with deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:3270-3278
[116] J'egou H,Douze M,Schmid C,et al.Aggregating local descriptors into a compact image representation[C]//IEEE Conference on Computer Vision and Pattern Recognition,2010:3304-3311
[117] Perronnin F,Liu Y,S'anchez J,et al.Large-scale image retrieval with compressed Fisher vectors [C] // IEEE Conference on Computer Vision and Pattern Recognition,2010:3384-3391
[118] Li F J,Tong W,Jin R,et al.An efficient key point quantization algorithm for large scale image retrieval[C]//ACM Workshop on Large-scale Multimedia Retrieval and Mining,2009:89-96
[119] Chu L Y,Wang S H,Zhang Y Y,et al.Graph density-based visual word vocabulary for image retrieval[C]//IEEE International Conference on Multimedia and Expo,2014:1-6
[120] Dong W,Wang Z,Charikar M,et al.Efficiently matching sets of features with random histograms[C]//ACM International Conference on Multimedia,2008:179-188

[121] Zhou W G, Yang M, Li H Q, et al. Towards codebook-free: Scalable cascaded hashing for mobile image search [J]. IEEE Transactions on Multimedia, 2014, 16(3): 601-611
[122] Zhang S L, Tian Q, Hua G, et al. Generating descriptive visual words and visual phrases for large-scale image applications [J]. IEEE Transactions on Image Processing, 2011, 20(9): 2664-2677
[123] Wang X Y, Yang M, Cour T, et al. Contextual weighting for vocabulary tree based image retrieval [C]// International Conference on Computer Vision, 2011: 209-216
[124] Liu Z, Li H Q, Zhou W G, et al. Embedding spatial context information into inverted file for large-scale image retrieval [C]// ACM International Conference on Multimedia, 2012: 199-208
[125] Liu Z, Li H Q, Zhou W G, et al. Contextual hashing for large-scale image search [J]. IEEE Transactions on Image Processing, 2014, 23(4): 1606-1614
[126] Chum O, Perdoch M, Matas J. Geometric min-hashing: Finding a (thick) needle in a haystack [C]// IEEE Conference on Computer Vision and Pattern Recognition, 2009: 17-24
[127] Bhat D N, Nayar S K. Ordinal measures for image correspondence [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(4): 415-423
[128] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories [J]. IEEE Conference on Computer Vision and Pattern Recognition, 2006, 2: 2169-2178
[129] Cao Y, Wang C H, Li Z W, et al. Spatial-bag-of-features [C]// IEEE Conference on Computer Vision and Pattern Recognition, 2010: 3352-3359
[130] Wu Z, Ke Q F, Sun J, et al. Scalable face image retrieval with identity-based quantization and multireferencer-eranking [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(10): 1991-2001
[131] Bentley J L. *K*-d trees for semi dynamic point sets [C]// Symposium on Computational Geometry, 1990: 187-197
[132] Silpa-Anan C, Hartley R. Localisation using an image map [C]// Australian Conference on Robotics and Automation, 2004
[133] Muja M, Lowe D G. Scalable nearest neighbor algorithms for high dimensional data [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(11): 2227-2240
[134] Zhou W G, Yang M, Wang X Y, et al. Scalable feature matching by dual cascaded scalar quantization for image retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 159-171
[135] Jain M, J'egou H, Gros P. Asymmetric hamming embedding: Taking the best of our bits for large scale image search [C]// ACM International Conference on Multimedia, 2011: 1441-1444
[136] Zhou W G, Li H Q, Lu Y J, et al. Visual word expansion and BSIFT verification for large-scale image search [J]. Multimedia Systems, 2013, 21(3): 245-254
[137] Xia Y, He K M, Wen F, et al. Joint inverted indexing [C]// IEEE International Conference on Computer Vision, 2013: 3416-3423
[138] J'egou H, Chum O. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening [C]// European Conference on Computer Vision, 2012: 774-787
[139] Zheng L, Wang S J, Zhou W G, et al. Bayes merging of multiple vocabularies for scalable image retrieval [C]// IEEE Conference on Computer Vision & Pattern Recognition, 2014: 1963-1970
[140] Indyk P, Motwani R. Approximate nearest neighbors: Towards removing the curse of dimensionality [C]// Annual ACM Symposium on Theory of Computing, 1998: 604-613
[141] Andoni A, Indyk P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions [C]// IEEE Symposium Foundations of Computer Science, 2006: 459-468
[142] Lv Q, Josephson W, Wang Z, et al. Multiprobelsh: Efficient indexing for high-dimensional similarity search [C]// International Conference on Very Large Data Bases, 2007: 950-961
[143] Wang J, Kumar S, Chang S F. Semi-supervised hashing for scalable image retrieval [C]// IEEE Conference on Computer Vision and Pattern Recognition, 2010: 3424-3431
[144] Gong Y C, Lazebnik S. Iterative quantization: A procrustean approach to learning binary codes [C]// IEEE Conference on Computer Vision and Pattern Recognition, 2011: 817-824
[145] Aiger D, Kokiopoulou E, Rivlin E. Random grids: Fast approximate nearest neighbors and range searching for image search [C]// IEEE International Conference on Computer Vision, 2013: 3471-3478
[146] Iwamura M, Sato T, Kise K. What is the most efficient way to select nearest neighbor candidates for fast approximate nearest neighbor search? [C]// IEEE International Conference on Computer Vision, 2013: 3532-3542
[147] Wang J D, Li S P. Query-driven iterated neighborhood graph search for large scale indexing [C]// ACM International Conference on Multimedia, 2012: 179-188
[148] Wang M, Zhou W G, Tian Q, et al. Linear distance preserving pseudo-supervised and unsupervised hashing [C]// ACM International Conference on Multimedia, 2016: 1257-1266
[149] Ge T Z, He K M, Ke Q F, et al. Optimized product quantization for approximate nearest neighbor search [C]// IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2946-2953
[150] Tuytelaars T, Schmid C. Vector quantizing feature space with a regular lattice [C]// International Conference on Computer Vision, 2007: 1-8
[151] Arandjelovic R, Zisserman A. All about VLAD [C]// IEEE Conference on Computer Vision and Pattern Recognition, 2013: 1578-1585
[152] Spyromitros-Xioufis E, Papadopoulos S, Kompatsiaris I, et al. A comprehensive study over VLAD and product quantization in for large-scale image retrieval [J]. IEEE Transactions on Multimedia, 2014, 16(6): 1713-1728
[153] J'egou H, Zisserman A. Triangulation embedding and

democratic aggregation for image search[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2014:3310-3317
[154] Gao Z N,Xue J R,Zhou W G,et al.Fast democratic aggregation and query fusion for image search[C]// ACM International Conference on Multimedia Retrieval,2016: 35-42
[155] Ge T Z,Ke Q F,Sun J.Sparse-coded features for image retrieval[C]// British Machine Vision Conference, 2013:132
[156] Liu Z,Li H Q,Zhou W G,et al.Uniforming residual vector distribution for distinctive image representation [J].IEEE Transactions on Circuits & Systems for Video Technology,2015,26(2):1-1
[157] Liu Z,Li H Q,Zhou W G,et al.Uniting keypoints:Local visual information fusion for large scale image search[J]. IEEE Transactions on Multimedia,2015,17(4):538-548
[158] Jaakkola T,Haussler D.Exploring generative model indiscriminative classifiers[C]// Proceedings of the 1998 Conference in Advances in Neural Information Processing Systems,1999:487-493
[159] Frasconi P.Learning with kernels and logical representations[J]. International Conference on Inductive Logic Programming,2007:1-3
[160] S'anchez J,Perronnin F,Mensink T,et al.Image classification with the Fisher vector:Theory and practice[J].International Journal of Computer Vision,2013,105(3): 222-245
[161] Duan L Y,Gao F,Chen J,et al.Compact descriptors for mobile visual search and MPEG CDVS standardization [C]// IEEE International Symposium on Circuits and Systems,2013:885-888
[162] Gong Y C,Wang L W,Guo R Q,et al.Multi-scale orderless pooling of deep convolutional activation features[C] // European Conference on Computer Vision, 2014: 392-407
[163] Yandex A B, Lempitsky V. Aggregating local deep features for image retrieval[C]// IEEE International Conference on Computer Vision,2015:1269-1277
[164] Baeza-Yates R A,Ribeiro-Neto B.Modern information retrieval[M]. New York: Addison-Wesley Longman Publishing Co.,Inc,1999
[165] Cai J J,Liu Q,Chen F,et al.Scalable image search with multiple index tables[C]// International Conference on Multimedia Retrieval,2014:407
[166] Zheng L,Wang S J,Tian Q.Coupled binary embedding for large-scale image retrieval[J].IEEE Transactions on Image Processing,2014,23(8):3368-3380
[167] Zhang X,Li Z W,Zhang L,et al.Efficient indexing for large scale visual search[C]// IEEE International Conference on Computer Vision,2009:1103-1110
[168] Silpa-Anan C, Hartley R. Optimized KD-trees for fast image descriptor matching[C]// IEEE Conference on Computer Vision and Pattern Recognition,2008:1-8
[169] Zheng L,Wang S J,Liu Z Q,et al.Fast image retrieval: Query pruning and early termination [J]. IEEE Transactions on Multimedia,2015,17(5):648-659
[170] Ji R R,Duan L Y,Chen J,et al.Learning to distribute vocabulary indexing for scalable visual search[J]. IEEE Transactions on Multimedia,2013,15(1):153-166
[171] Heo J P,Lee Y,He J F,et al.Spherical hashing[C]// IEEE Conference on Computer Vision and Pattern Recognition,2012:2957-2964
[172] Tang J H,Li Z C,Wang M,et al.Neighborhood discriminant hashing for large-scale image retrieval[J]. IEEE Transactions on Image Processing, 2015, 24(9): 2827-2840
[173] Wu L, Zhao K, Lu H T, et al. Distance preserving marginal hashing for image retrieval[C]// IEEE International Conference on Multimedia and Expo,2015:1-6
[174] Jiang K,Que Q C,Kulis B.Revisiting kernelized locality sensitive hashing for improved large-scale image retrieval [C]// IEEE Conference on Computer Vision and Pattern Recognition,2015:4933-4941
[175] Liu H M,Wang R P,Shan S G,et al.Deep supervised hashing for fast image retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2064-2072
[176] Datar M,Immorlica N,Indyk P,et al.Locality sensitive hashing scheme based on p-stable distributions[C]// Proceedings of the Twentieth Annual Symposium on Computational Geometry,2004:253-262
[177] Avrithis Y,Tolias G,Kalantidis Y.Feature map hashing: Sub-linear indexing of appearance and global geometry [C]// International Conference on Multimedia, 2010: 231-240
[178] Tolias G,Kalantidis Y,Avrithis Y,et al.Towards large-scale geometry indexing by feature selection[J]. Computer Vision and Image Understanding, 2014, 120 (2):31-45
[179] J'egou H, Douze M, Schmid C. Packing bag-of-features [C]// International Conference on Computer Vision, 2009:2357-2364
[180] Chum O,Philbin J,Isard M,et al.Scalable near identical image and shot detection[C]// ACM International Conference on Image and Video Retrieval,2007:549-556
[181] Lin Z,Brandt J.A local bag-of-features model for large scale object retrieval[C]// European Conference on Computer Vision,2010:294-308
[182] J'egou H,Schmid C,Harzalla H,et al.Accurate image search using the contextual dissimilarity measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2010,32(1):2-11
[183] Qin D F, Wengert C, Van Gool L. Query adaptive similarity for large scale object retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2013:1610-1617
[184] Donoser M, Bischof H. Diffusion processes for retrieval revisited[C]// IEEE Conference on Computer Vision and Pattern Recognition,2013:1320-1327
[185] Zheng L,Wang S J,Liu Z,et al.$L(p)$-norm IDF for large scale image search[C]// IEEE Conference on Computer Vision and Pattern Recognition,2013:1626-1633
[186] Zheng L,Wang S J,Tian Q.$L(p)$-norm IDF for scalable image retrieval[J].IEEE Transactions on Image Processing,2014,23(8):3604-3617

[187] Shen X H,Lin Z,Brandt J,et al.Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2012: 3013-3020

[188] Xie H T, Gao K, Zhang Y D, et al. Pairwise weak geometric consistency for large scale image search[C]// ACM International Conference on Multimedia Retrieval, 2011:42

[189] Katz S M.Distribution of content words and phrases in text and language modeling[J].Natural Language Engineering,1996,2(1):15-59

[190] J'egou H,Douze M,Schmid C.On the burstiness of visual elements[C]// IEEE Conference on Computer Vision and Pattern Recognition,2009:1169-1176

[191] Shi M J,Avrithis Y,J'egou H.Early burst detection for memory-efficient image retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2015: 605-613

[192] Bai S,Bai X.Sparse contextual activation for efficient visual re-ranking[J].IEEE Transactions on Image Processing,2016,25(3):1056-1069

[193] Yang F,Matei B,Davis L S.Re-ranking by multi-feature fusion with diffusion for image retrieval[C]// IEEE Winter Conference on Applications of Computer Vision, 2015:572-579

[194] Li X C, Larson M, Hanjalic A. Pairwise geometric matching for large-scale object retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2015:5153-5161

[195] Kuo Y H,Chen K T,Chiang C H,et al.Query expansion for hash-based image object retrieval[C]// ACM International Conference on Multimedia,2009:65-74

[196] Chum O, Matas J. Matching with PROSAC-progressive sample consensus[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2005: 220-226

[197] Avrithis Y,Tolias G.Hough pyramid matching:Speeded-up geometry re-ranking for large scale image retrieval [J].International Journal of Computer Vision,2014,107 (1):1-19

[198] Fischler M A,Bolles R C.Random sample consensus:A paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM,1981,24(6):381-395

[199] Jiang Y G, Jiang Y, Wang J. VCDB: A large-scale database for partial copy detection in videos[C]// European Conference on Computer Vision,2014:357-371

[200] Grauman K, Darrell T. The pyramid match kernel: Discriminative classification with sets of image features[C] // IEEE International Conference on Computer Vision, 2005,2:1458-1465

[201] Zhou W G,Li H Q,Lu Y J,et al.SIFT match verification by geometric coding for large-scale partial-duplicate web image search[J].ACM Transactions on Multimedia Computing,Communications,and Applications,2013,9(1):4

[202] Chu L Y,Jiang S Q,Wang S H,et al.Robust spatial consistency graph model for partial duplicate image retrieval [J]. IEEE Transactions on Multimedia, 2013, 15(8): 1982-1996

[203] Xie L X, Tian Q, Zhou W G, et al. Fast and accurate near-duplicate image search with affinity propagation on the image web[J]. Computer Vision and Image Understanding,2014,124:31-41

[204] Kleinberg J M.Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632

[205] Xie L X,Tian Q,Zhou W G,et al.Heterogeneous graph propagation for large-scale web image search[J]. IEEE Transactions on Image Processing, 2015, 24(11): 4287-4298

[206] Xie H,Zhang Y,Tan J,et al.Contextual query expansion for image retrieval [J]. IEEE Transactions on Multimedia,2014,16(4):1104-1114

[207] Tao D C,Tang X O.Random sampling based SVM for relevance feedback image retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2004: 647-652

[208] Tao D C, Tang X O, Li X L, et al. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006,28(7):1088-1099

[209] Hoi S C H,Jin R,Zhu J K,et al.Semi-supervised SVM batch mode active learning for image retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition,2008:1-7

[210] Arevalillo-Herr'aez M, Ferri F J. An improved distance based relevance feedback strategy for image retrieval[J]. Image and Vision Computing,2013,31(10):704-713

[211] Rabinovich E, Rom O, Kurland O. Utilizing relevance feedback in fusion-based retrieval[C]// International ACM SIGIR Conference on Research & Development in Information Retrieval,2014:313-322

[212] Wang X Y,Li Y W,Yang H Y,et al.An image retrieval scheme with relevance feedback using feature reconstruction and SVM reclassification[J].Neurocomputing,2014, 127:214-230

[213] Tieu K, Viola P. Boosting image retrieval [J]. International Journal of Computer Vision, 2004, 56(1/2):17-36

[214] Yu J,Tao D,Wang M,et al.Learning to rank using user clicks and visual features for image retrieval[J]. IEEE Transactions on Cybernetics,2015,45(4):767-779

[215] Zhou X S,Huang T S.Relevance feedback in image retrieval:A comprehensive review[J].Multimedia Systems, 2003,8(6):536-544

[216] Patil P B, Kokare M B. Relevance feedback in content based image retrieval: A review[J]. Journal of Applied Computer Science & Mathematics,2011,5(10):41-47

[217] Fagin R, Kumar R, Sivakumar D. Efficient similarity search and classification via rank aggregation[C]// ACM SIGMOD International Conference on Management of Data,2003:301-312

[218] Page L, Brin S, Motwani R, et al. The pagerank citation ranking: Bringing order to the web[J]. Stanford Digital

Libraries Working Paper,1998,9(1):1-14
[219] Ye G N,Liu D,Jhuo I H,et al.Robust late fusion with rank minimization[C]//IEEE Conference on Computer Vision and Pattern Recognition,2012:3021-3028
[220] Romberg S,Pueyo L G,Lienhart R,et al.Scalable logo recognition in real-world images[C]//ACM International Conference on Multimedia Retrieval,2011:25
[221] Wang S,Jiang S Q.INSTRE:A new benchmark for instance-level object retrieval and recognition[J].ACM Transactions on Multimedia Computing,Communications,and Applications,2015,11(3),DOI:10.1145/2700292
[222] Chandrasekhar V R,Chen D M,Tsai S S,et al.The Stanford mobile visual search data set[C]//ACM Conference on Multimedia Systems,2011:117-122
[223] Tian X M,Lu Y J,Yang L J,et al.Learning to judge image search results[C]//ACM International Conference on Multimedia,2011:363-372
[224] Tian X M,Jia Q H,Mei T.Query difficulty estimation for image search with query reconstruction error[J].IEEE Transactions on Multimedia,2015,17(1):79-91
[225] He K M,Zhang X Y,Ren S Q,et al.Spatial pyramid pooling indeep convolutional networks for visual recognition[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2015,37(9):1904-1916

Recent advance in content-based image retrieval:A literature survey

ZHOU Wengang[1] LI Houqiang[1] TIAN Qi[2]
1 School of Information Science and Technology,University of Science and Technology of China,Hefei 230026
2 The Department of Computer Science,University of Texas at San Antonio,San Antonio,TX 78249,USA

Abstract The explosive increase and ubiquitous accessibility of visual data on the Web have led to the prosperity of research activity in image search or retrieval.With the ignorance of visual content as a ranking clue,methods with text search techniques for visual retrieval may suffer inconsistency between the text words and the visual content.Content-based image retrieval(CBIR),which makes use of the representation of visual content to identify relevant images,has attracted sustained attention in recent two decades.Such a problem is challenging due to the intention gap and the semantic gap problems.Numerous techniques have been developed for content-based image retrieval in the last decade.The purpose of this paper is to categorize and evaluate those algorithms proposed during the period of 2003 to 2016.We conclude with several promising directions for future research.

Key words content-based image retrieval;visual representation;indexing;similarity measurement;spatial context;search re-ranking