中国科学: 信息科学　2013 年　第 43 卷　第 12 期: 1641–1653
www.scichina.com　info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS
评　述　中国科学院学部　科学与技术前沿论坛　信号与信息处理专刊

大规模互联网图像检索与模式挖掘

张磊

微软亚洲互联网工程院, 北京 100080
E-mail: leizhang@microsoft.com

收稿日期: 2013–05–28; 接受日期: 2013–09–29

摘要　在互联网时代, 爆炸式增长的数字图像不仅给图像检索带来巨大的技术挑战, 同时也带来了很多机遇和研究问题的新思路. 本文简单回顾了图像检索的三个阶段的研究历史, 以及在此过程中数据量的增多给图像检索带来的影响, 并对作为关键问题的特征提取方面的研究进行了深入的分析. 本文尤其指出视觉模式挖掘是寻找中层特征表示并缩小语义鸿沟的重要研究方向, 并根据视觉模式的表征粒度将其分为五种类别分别进行了介绍, 从中可以看到大数据对于视觉模式挖掘的重要作用.

关键词　图像检索　视觉模式　模式挖掘　内容分析　大数据

1　引言

图像检索是管理个人数字照片和快速获取互联网图像的重要技术. 随着数码相机和可拍照智能手机的广泛普及, 以及互联网上数量众多的社区网络用户的参与, 无论是个人照片还是互联网图像的数量都在呈爆炸式增长. 仅以 Facebook 为例, 2011 年 2 月该网站累积有 600 亿用户上传图像 [1)], 2013 年 1 月该数量就增加至 2400 亿之多 [2)]. 图像数量的急剧增多, 不仅给存储系统带来巨大的困难, 对于如何快速访问和查找用户所需要的图像也带来前所未有的挑战.

图像检索中, 用户可以通过多种方式表达其查询意图, 如关键词、示例图像、线条图等. 不管是哪种形式的查询表示, 一个图像检索系统都需要对图像进行分析和理解, 从而才能在图像库中找到匹配用户查询意图的图像, 因此图像内容分析一直是图像检索中的基础研究问题. 在图像内容分析的研究中, 研究人员关注的焦点是如何从图像中提取有效的特征和表示. 然而即使经过数十年的研究, 该问题仍然没有被很好地解决. 其困难在于众所周知的语义鸿沟, 即从图像中提取出的低层特征和图像所表示的高层语义概念之间存在的明显差距, 很难靠简单的特征提取和函数映射来解决. 因此, 图像检索既有巨大的应用需求, 又困难重重, 其核心问题同时也是计算机视觉中物体识别的关键问题, 长期受到众多研究人员的关注.

为了解决图像内容分析中的语义鸿沟问题, 研究人员近年来开始尝试寻找介于低层特征和高层语义之间的中间表示, 试图通过这样的中间表示来缩小这个鸿沟. 这些中间层表示在图像中往往对应于

1) http://thenextweb.com/socialmedia/2011/08/05/flickr-hits-6-billion-total-photos-but-facebook-does-that-every-2-months/

2) http://newsroom.fb.comp/News/562/Introducing-Graph-Search-Beta

引用格式: 张磊. 大规模互联网图像检索与模式挖掘. 中国科学: 信息科学, 2013, 43: 1641–1653, doi: 10.1360/N112013-00058

一些可重复的、结构化的、对应于物体部件的视觉模式单元, 这些视觉模式单元相比于基于统计的低层特征来说, 更接近于语义表示并且具有更好的鲁棒性.

大规模互联网图像的出现, 给图像检索不仅带来很多挑战, 也带来很多机遇. 首先, 数量巨大的网络图像给研究工作提供了用之不尽的训练数据, 对于图像内容分析、视觉概念建模、物体识别等问题有着极大的帮助; 其次, 与个人数字照片不同, 互联网图像是和网页及用户紧密关联在一起的, 具有丰富的弱标注信息, 如网页中的上下文文本信息和用户在社区网络中提供的简单标注, 这些弱标注信息给图像内容分析和图像检索提供了极为有价值的语义信息; 再次, 网络图像因为数量巨大并有冗余, 在数据中呈现出丰富的可重复视觉模式, 比如重复的图像、物体、结构化视觉单元等, 正是因为这些冗余信息, 才使得数据挖掘和机器学习算法能够从数据中归纳学习出具有推广能力的结构化模式单元.

本文将主要讨论互联网图像检索中的挑战和大规模数据所带来的机遇, 以及视觉模式挖掘对于解决图像内容分析的意义和主要研究进展. 主要内容包括: 1) 图像检索研究的回顾; 2) 大规模数据对图像检索带来的巨大影响; 3) 视觉模式挖掘对于图像内容分析的意义; 4) 视觉模式挖掘的主要研究现状和发展趋势.

2 图像检索研究回顾

2.1 图像检索研究的三个阶段

图像检索可以追溯到 1970 年代, 其研究过程可以大体分为以下三个阶段 [1]:

基于文本检索的阶段 (1970 ∼ 1990). 这一阶段的检索系统通常把图像检索的问题转换成传统的文本检索问题, 这样可以借助许多相对较为成熟的数据库技术来解决. 即首先对图像用文本进行标注, 然后用基于文本的数据库管理系统 (DBMS) 来进行图像检索. 然而, 由于图像标注很难用自动的方法完成, 大多图像检索系统需要依靠人工进行图像标注. 这个方法对于小规模图像库尚可操作, 但是对于大规模图像库 (如包括数万以上图像), 基于人工标注的方法就难以奏效了. 除了人工标注的工作量过大, 另外图像标注还有很强的主观性和不精确性, 这是由于图像内容的丰富性和个人感知的主观性之间的矛盾引起的, 是一个难以克服的根本问题.

基于图像内容检索的阶段 (1990 ∼ 2000). 从 1990 年代开始, 由于数码相机和数字扫描设备的出现, 数字图像的数量开始急剧增多, 用户开始希望能够基于图像内容本身进行检索和管理, 而早期的基于文本的图像检索方法显然无法满足这些需求. 1992 年, 美国自然科学基金组织了一个关于视觉信息管理系统 (visual information management systems) 的研讨会 [2] 以确定图像数据库管理系统方面新的研究方向. 从此以后, 来自计算机视觉、数据库、人机交互、信息检索等研究领域的研究人员开始加入基于内容的图像检索的研究, 提出许多图像特征提取、索引、用户交互等算法, 并研制出许多商业及演示系统. 这一阶段遇到的最大的困难即是所谓的语义鸿沟问题. 用户希望在语义层面查找相似图像, 而计算机只能通过处理图像中的原始像素数据, 或是处理从像素中提取出的低层视觉特征来定义图像之间的相似度, 这些特征往往是对图像的色彩、纹理、形状的统计, 无法真正反映图像的语义. 为了弥补这个语义鸿沟, 研究人员提出相关反馈的方法 [3], 希望用户能对检索结果提供反馈信息, 从而再次检索以得到更好的结果. 这个方法的主要问题是反馈操作过程繁琐且结果改进有限, 用户往往不愿意配合.

网络图像检索的阶段 (2000 ∼ 现在). 从 2000 年开始, 互联网和数字成像设备开始普及, 导致网络图像数量剧增. 2001 年, 谷歌公司首次发布了它的图像搜索引擎. 该系统利用网络图像所在网页的

上下文 (如文件名、URL、标题、离图像较近的文字等) 作为图像的描述文字, 索引了 2.5 亿网络图像供用户查询. 虽然这是纯粹的基于文本的图像检索系统, 但是检索结果通常能够满足用户需求. 商用图像搜索系统的成功, 使得研究人员深受鼓舞, 并且开始研究大规模网络图像检索中的一系列问题 [1]. 比如: 1) 如何对图像自动进行标注; 2) 如何创建高效的索引系统以检索更大规模的图像库; 3) 如何隐式地收集用户的反馈信息来改进图像搜索引擎的性能; 4) 如何支持更多的查询模态来帮助用户更好地表达查询意图. 尤其后两个问题与人机交互和用户意图理解相关, 虽然这些问题在图像检索研究的早期研究阶段就受到关注, 但是近几年来由于大规模商用图像检索系统的出现和新的交互设备 (如触摸屏) 的普及, 关于这些问题的研究也有了显著的进展, 例如文献 [4] 如何有效利用用户查询日志, 文献 [5, 6] 通过用户画的彩色布局图或线条图来理解查询意图并进行检索; 文献 [7] 提出通过图文结合的方式给用户进行查询推荐. 同时, 在这一阶段中, 低层特性的研究开始转向具有局部不变性的特征, 并用来解决重复图像检测和物体识别的问题.

另一个值得关注的方面是, 近年来社区网络和移动计算给图像检索带来了更多的新数据、新问题和新应用. 如 Flickr 和 Facebook 两个网站积累了数十亿甚至数千亿的图像, 而且很多图像带有用户提供的简单标注, 这给图像搜索、标注和排序提供了非常有价值的信息. 另外, 装备有摄像头的智能手机已经广为普及, 通过手机拍照来查询并识别物理世界中的物体也变得非常有用.

2.2 大数据的兴起

在过去二十年的图像检索的研究过程中, 我们可以很明显地看到图像库规模在不断增大. 在 2000 年以前, 大多的研究工作所使用的图像库仅包含数千至数万幅图像. 如 IBM 的 QBIC 系统仅采用了 1,000 幅图像 [8], 加州大学圣芭芭拉分校的 Netra 系统采用了 2,500 幅 Corel 图像 [9], 哥伦比亚大学的 VisualSEEk 系统索引了 12,000 幅图像 [10], 其 WebSEEk 系统大概是 2000 年前最大的系统, 用基于文本的方法索引了 513,323 幅网络图像 [11], 由于检索效率的问题, 基于内容的检索仅用于改进基于文本的检索结果.

2000 年以前的研究中所使用的图像数量少有两方面的原因, 一是图像获取还比较困难, 二是高维空间索引的问题还没有被很好的解决. 2000 年以后, 互联网的普及使得图像获取不再是一个阻碍研究的问题. 另一方面, 2000 年左右的几个重要的技术突破对图像检索起了极大的推动作用. 这些技术突破包括 [1]:

特征. 1999 年, Lowe[12] 提出了具有尺度不变性的 SIFT 局部特征检测和描述方法, 基本解决了图像之间局部点匹配的问题, 在重复图像检索和物体识别中得到成功的应用. 因此, 2000 年以后的特征提取方面的研究也明显地由全局特征转向局部特征.

索引. 1999 年, Gionis 等 [13] 提出了局部敏感哈希 (LSH) 的方法, 用于解决高维空间中快速 K- 近邻查找的问题. 该方法及其各种改进算法使得在图像检索系统中高效地索引上百万甚至于上亿的图像成为可能.

系统. 2001 年, 谷歌公司发布了索引有 2.5 亿图像的搜索系统, 成为图像检索在商业应用中的重要里程碑. 尽管这是一个基于文本的图像检索系统, 但是由于它索引了上亿图像并有效地展示了网络图像上下文的重要作用, 因此吸引了很多研究人员开始关注并研究自动图像标注和高维索引等问题.

除此之外, 另一个重要的技术突破是统计学习理论及其应用算法支持向量机 (SVM)[14], 1995 年后在很多和分类相关的应用中取得令人瞩目的成功, 并对机器学习领域产生了深远的影响. 在 2000 年以后的研究工作中, SVM 及其他分类算法被广为应用于相关反馈、图像标注、视觉概念表示、物体分

类等研究问题中.

由于这些技术上的突破, 在 2000 年后的研究工作中, 图像库的规模持续增大. 如 Wang 等 [15] 开发的 SIMPLIcity 系统索引了 20 万幅图像, Quack 等 [16] 开发的 Cortina 系统使用了 300 万幅图像, Wang 等 [17] 以及 Li 等 [18] 采用了 240 万幅图像来解决图像标注问题; 为了验证数据规模对于图像标注和物体识别问题的影响, Torralba 等 [19] 收集了 8 千万幅图像, 并采用简单的 K- 近邻搜索和投票的方法, 取得了很好的结果; 为了进一步利用大数据, Wang 等 [20] 把图像库的规模扩大到 20 亿, 并采用重复图像检索的方法对网络上重复度较高的图像进行准确的标注.

图像数量的增加对图像检索的研究产生了深刻的影响. 互联网上大量增加的图像数据, 不仅给图像检索带来了许多索引和系统方面的挑战, 同时也带来了很多机遇. 人工智能领域中有很多用传统方法难以解决的问题, 因为互联网提供的海量训练数据, 使得求解这些问题成为可能. 图像检索和图像理解问题也正在受益于这些超大规模的网络图像数据.

2.3 视觉特征研究

图像检索中的一个核心问题就是如何度量两个图像之间的相似度. 这个问题对于人来说并不算难, 人可以很容易地判断出两个图像是否相似以及在哪些方面相似. 然而对于计算机来说, 这个问题成了迄今为止仍未能解决的难题. 这是由于我们对人类大脑的认知机理的认识还处于非常初级的阶段, 虽然神经认知科学的研究告诉我们, 大脑处理和识别外部视觉信号的过程是一个自底向上逐层抽象的过程, 但是对于大脑如何对视觉信号进行抽象和推广我们所知甚少. 因此, 虽然计算机擅长于快速的数值计算, 但是对于如何从图像数据中提取出对人来说有语义的信息仍然是极端困难的问题.

在 2000 年以前, 研究人员主要关注如何从整个图像中提取一些统计特征, 从而可以比较两个图像的相似度. 虽然这个相似度不是语义上的相似度, 但是当用户提供一个如日出或建筑等色彩或纹理特征非常明显的查询图像时, 检索系统还是可以返回不错的结果的. 这些特征主要是对颜色、纹理和形状方面的统计, 如颜色直方图、颜色相关直方图、基于小波变换系数的纹理统计、形状的统计描述子等. 这类特征的最主要的问题是和语义概念之间存在巨大的差距, 即语义鸿沟, 因此其应用范围非常有限. 关于这个时期特征提取研究的综述可参见文献 [21].

2000 年以后, 受到 SIFT 特征在很多应用上取得的成功的鼓舞, 特征提取方面的研究明显地由全局特征转向局部特征. SIFT 特征的基本原理是检测图像信号中的局部不变量, 从而得到较强的平移、缩放和旋转变换不变性, 以及一定程度的透视和光照变化的不变性. 该特征 (包括检测子和描述子) 对于计算机视觉中很多问题 (如图像分类、检索、物体识别、三维重建等) 都起到了巨大的推动作用, 也促使研究人员进一步研究更为有效的特征表示用于图像分类和物体识别.

SIFT 特征描述子是对图像局部区域中的梯度信号 (即边缘) 的统计. 采用梯度信号的好处是它对图像的光照变化不敏感, 另外图像中的边缘信息对于视觉认知来说也更为重要. 由于 SIFT 描述子是为了解决图像匹配问题而提出的, 对于匹配的要求较为严格, 为了取得更好的鲁棒性, Dalal 等 [22] 还提出了梯度直方图 (HOG) 特征, 在行人检测 [22] 及图像检索 [23] 中取得了很好的结果.

采用类似的思路, Ahonen 等 [24] 还提出了 local binary pattern (LBP) [25] 特征, 采用局部区域中周边像素值和中心像素值的差值并做二值化来对局部区域进行编码, 从而刻画局部区域的反差和空间关系信息, 在纹理图像分析、人脸识别和人脸图像检索等应用中取得了很好的效果.

SIFT 以及其他局部关键点特征的主要问题是对于物体检测和语义表示来说区分能力不足. SIFT 本质上是用于解决图像匹配 (尤其是立体视觉) 问题的, 无论是特征检测还是特征描述都是为了匹配

相同的物体, 而不是面向图像内容和语义分析的. 另外单一的局部关键点也很难表示明确的语义. 为了使其具备更强的区分能力, 很多研究考虑了如何利用 SIFT 特征点的组合来增加其区分性. 这类方法对于刚体类物体 (如建筑) 具有显著的效果. 但是对于动物、植物等非刚体物体其推广能力非常有限. 这是因为依赖 SIFT 特征的表示一般具有较高的匹配准确率, 但无法解决相似意义上的模式匹配.

近些年来, 越来越多的研究人员意识到以前研究工作的局限性. 基于统计的全局特征比较鲁棒, 但是不能刻画出图像的语义; 基于局部关键点的特征可以很好的用来检测重复结构, 但是推广能力非常有限. 因此, 很多研究开始寻找视觉特征的中层表示, 以弥补底层特征和高层语义间的差距. 这些中间层表示往往是由低层特征组合而成, 通过引入局部特征点的位置信息以反映局部结构; 或是采用机器学习的方法从数据中学习出在语义上比较一致并具有一定的结构信息的视觉模式, 这成为近年来图像检索、计算机视觉和机器学习的关键问题之一.

3 视觉模式的特点

视觉模式是指在大规模图像数据中存在的可重复的、结构化的、对应于某些物体部件的视觉模式单元, 这些视觉模式单元相比于基于统计的低层特征来说, 更接近于语义表示并且有较强的鲁棒性. 对于理想的视觉模式, 我们尝试将其特点总结如下:

重复性. 视觉模式在图像数据中多次出现并具有一定的统计规律. 模式的重复性是自然界中固有的一个性质, 比如生物在进化和生长过程中是靠不断复制以及少量变异来进行的, 人造物体更是具有大量的重复结构和部件, 再加上人们拍照时会多次拍摄到同一物体, 因此在大规模图像数据中必然会有大量的重复视觉模式.

不变性. 理想情况下, 视觉模式对平移、尺度、旋转、透视等变换应具有一定不变性和鲁棒性. 但在实际应用中, 即使仅具有有限不变性的视觉模式也仍会有很高的应用价值. 比如 SIFT 特征仅具备平移、尺度、旋转不变性, 和有限的透视变换不变性; 重复图像也可以看作是一种简单的重复视觉模式, 具有尺度变换不变性和有限的旋转平移不变性. 即使如此大规模重复图像检测和聚类在图像检索和图像知识库构建中仍具有重要的应用价值 [26].

结构性. 视觉模式通常反映图像数据中的结构信息, 如视觉数据中的空间结构, 一般难以有效地显式表达. 视觉模式的挖掘通常需要和机器学习方法结合, 从数据中学习和挖掘结构化信息.

完备性. 这个特点是指一个视觉模式库中的模式数量应该足够多, 这样才可以用于分析和解释新的图像. 因此, 从大规模数据中挖掘出尽可能多的视觉模式对于图像内容分析和图像检索有着重要的意义. 尽管大规模图像数据中所包含的视觉模式数量庞大甚至是无穷的, 对于任何单一图像来说, 所包含的视觉模式往往是有限的和稀疏的, 在图像分析时通常可以施加稀疏约束来求解.

需要指出的是, 关于视觉模式的研究对图像检索有着重要的意义, 有助于从图像中提取更接近语义的特征, 因而可以有效地缩小语义鸿沟. 考虑到近年来的研究趋势, 本文将视觉模式挖掘作为重要的研究方向特别予以分析和展望.

4 视觉模式挖掘及应用

近几年来, 图像检索和计算机视觉中涌现出大量的和模式挖掘相关的工作. 根据视觉模式的表征粒度, 我们可以将其分为五个类别: 重复图像, 基于全图统计的类别模式, 基于局部关键点的视觉单词

组, 基于局部结构的视觉模式, 以及基于隐层表示的视觉模式. 以下我们分别予以介绍.

4.1 重复图像

重复图像检测是图像检索中的一个子问题, 自 2000 年以后由于个人数字照片和网络图像的增多而受到关注 [27]. 对于个人照片, 用户希望能够合并重复照片以方便管理; 对于网络图像, 重复图像检测可以用于版权保护; 对于图像搜索引擎, 则希望通过重复图像检测来提高图像库的质量.

重复图像可以被看作是一种最为简单的视觉模式, 虽然在很多应用中重复图像被看作是冗余数据而需要去除, 但是这种数据冗余性对图像分析却有很大的帮助, 如图像标注 [20]、图像知识库构建 [26]、三维重建 [28] 等.

在文献 [20] 中, Wang 等使用了 20 亿幅网络图像, 采用重复图像检索的方法来解决图像标注的问题. 对于一个待标注图像, 该方法将此图像作为查询图像从 20 亿幅图像中查找与其重复的图像, 如果系统能够返回足够多 (如 10 张以上) 的图像, 则系统对返回结果图像所关联的网页文本进行分析, 提取出现频率较高的关键词, 作为对查询图像的标注结果 (图例见文献 [20] 中图 1); 如果系统不能查找到重复图像, 则拒绝对此图像进行标注. 采用这种方法标注的结果具有非常高的准确度, 并能产生具体的人名和地名等专有名词, 这是传统的图像标注方法难以做到的. 在这个工作中, 作者对 20 亿幅网络图像进行了采样统计, 发现对于 8.1% 的图像在库中可以找到 10 张以上的重复图像, 这些图像通常是互联网上用户比较感兴趣的图像, 如关于名人、产品、标志性建筑、卡通、商标等的图像, 对这些图像的标注结果非常有助于提高图像搜索排序的相关度.

文献 [20] 中的工作主要是利用了重复图像检索的方法, 即给定一个查询图像, 在大规模图像库中查找其重复图像. 即使对数十亿的图像库, 通过高效的索引也可以实时地查找到所需图像. 但是在有些应用中, 需要检测出图像库中所有的互相重复的图像, 比如从众多图像中找到所有的重复图像后, 可以成批地进行图像标注以帮助改进图像搜索引擎 [20], 或者针对建筑图像进行建筑物的三维重建 [28], 或者将重复图像和大规模知识图谱 (knowledge graph) 进行匹配以建立大规模图像知识库 [26]. 检测图像库中所有重复图像的问题本质上是一个聚类问题, 相比之下, 这个问题的复杂度远高于重复图像检索的问题, 尤其是对超大规模 (如数十亿) 的图像库来说其时间代价非常之高.

主要的重复图像检测方法均是基于局部关键点的重复度来判断两幅图像是否重复的, 除了简单的图像缩放和压缩格式变换, 这种方法还可以检测局部重复的图像. 在文献 [28] 中, Chum 等利用 MinHash 的方法, 将每一图像的局部关键点集合通过随机算法转化成 MinHash 代码, 并进一步通过随机算法将 MinHash 代码组合成多个 Sketch. 然后该方法通过检测不同图像间 Sketch 的冲突来确定候选的重复图像 [28]. 这个方法可以有效地将局部重复的图像聚类在一起, 可以为建筑物三维重建提供丰富的数据. 但是文献 [28] 中的方法仅在数百万的图像库上进行了验证.

为了解决超大规模的图像聚类问题, Wang 等 [29] 采用了较为简单高效的全局特征, 在 20 亿幅图像中检测所有的重复图像 [26]. 聚类算法采用了分而治之的策略, 首先利用由全局特征生成哈希码将图像划分到若干小空间, 然后在小空间中进行图像聚类以降低复杂度, 因为空间划分会将一些聚类分隔开, 因此对前一步获取的聚类需要再进行合并. 该算法在具有上千个节点的并行计算平台上可以在一天的时间完成所有聚类的计算. 如前所述, 这些重复图像通常都是互联网上用户较为感兴趣的图像, 在文献 [26] 中作者进一步利用文献 [20] 中的图像标注方法对所有的重复图像组进行自动标注, 并将标注结果与大规模的知识图谱进行匹配, 从而建立大规模的图像知识库; 迄今为止, 已经收集了关于 52 万语义概念的 2.35 亿图像.

这个工作表明, 即使对于最简单的视觉模式 (如重复图像), 如果能从大规模数据中挖掘出足够多的模式, 并且和文本信息结合以确认有明确语义的模式, 仍是具有很高的应用价值的.

4.2 基于全图统计的类别模式

基于全图统计的类别模式, 是指采用分类的方法对图像类别建立分类器, 通过这些分类器可以对新图像进行分类, 以得到图像的类别信息. 这类方法相当于直接采用机器学习的方法学习从低层特征到高层语义的映射函数, 通常很难奏效. 但是在面向快速检索 [30], 或是训练图像经过很好的筛选的情况下 [31], 仍是可以产生很好的结果的.

在文献 [30] 中, Torresani 等提出 Classemes 的方法. 作者从多媒体领域中的一个大规模概念集 LSCOM (large scale concept ontology for multimedia) [32] 中选取了 2,659 个概念, 对每个概念从必应图像搜索引擎收集其返回的前 150 幅图像作为该概念的训练图像. 然后作者从图像中提取多种低层特征, 对每个概念训练分类器, 从而获取到和 2,659 个概念对应的 2,659 个分类器. 任何一幅新图像都可以被这 2,659 个分类器分类并转换为一个 2,659 维的向量, 这相当于把图像转换到了 2,659 维的语义空间, 在这个空间中可以进行更接近语义相似度的检索. 为了进一步提高检索效率, 作者还提出将每个分类器输出的结果二值化, 得到一个 2,659 位的二进制码, 这样的二进制码可以极大地提高图像检索效率并降低系统的存储开销. 在作者的后续工作中 [32], 将分类问题分为两步, 首先将图像特征通过投影和量化变换为二进制码, 然后对二进制码训练分类器, 这两步在一个目标函数中同时求解, 从而使得二进制码的学习可以直接有助于分类准确率的提高.

在文献 [31] 中, Tsai 等构造了大规模的 Visual Synset (视觉同义词组) 用于图像内容分析. 作者利用搜索引擎收集了关于 30 万关键词的 2 千万幅网络图像, 这些关键词均为搜索引擎中的高频搜索词, 对于每个关键词, 可以从搜索引擎收集到最多 1000 个图像, 并对这些图像进行基于视觉相似度的聚类, 从而得到视觉和语义上都比较相似的数量达数百万之多的图像组, 每个图像组被称为一个 Visual Synset (图例见文献 [31] 中图 1). 这些 Visual Synset 实质上可被看作具有显著视觉特征的基本语义单元, 这些图像组被用来训练多类线性 SVM 分类器, 以获得推广能力对新图像进行内容分析, 如图像分类和标注等. 由于类别数 (数百万) 和图像数量 (数千万) 都非常庞大, 这个研究工作利用了谷歌公司的 MapReduce 分布式计算平台进行并行计算和数据处理. 这个工作实质上是在海量的互联网图像库中 (通过搜索引擎和聚类算法) 挖掘出视觉特征显著并且和语义有较强关联的可视模式 (visual pattern), 并通过机器学习中的分类算法使其具有一定的推广能力. 这些具有推广能力的可视模式对于图像内容分析具有非常重要的意义.

采用全局视觉特征来表示图像, 由分类器学习所得的仍然是较粗略的全图为主的模式. 但是在大规模的问题中, 为了效率而采用全局特征, 仍是可以在应用问题中取得很好的结果的.

4.3 基于局部关键点的视觉词组

基于局部关键点特征如 SIFT 的图像表示, 对物体的平移、缩放、旋转、光照、遮挡等变化具有很好的不变性, 因而在图像检索中被广为采用. 对于局部关键点特征, 词袋模型 (bag of visual words) 和倒排索引 (inverted index) 是最为广泛使用的方法 [33]. 但是词袋模型最大的问题是丢掉了特征点在图像中的位置信息, 缺乏足够的区分力, 因而会导致错误的图像匹配. 这是因为单个视觉单词仅反映了图像中的局部信号, 并不对应具体的语义. 为了解决这个问题, 有不少工作研究如何将特征点组合成视觉词组 (visual phrase) 来增加其区分性.

Yuan 等 [34] 的工作是关于这个问题比较早的研究工作. 在这个工作中, 视觉词组被定义为图像中位置相关并在多个图像中共同出现的多个视觉单词. 作者采用了数据挖掘领域经典的频繁项集挖掘 (frequent itemset mining) 算法 [35], 从 435 幅人脸图像和 123 幅汽车图像挖掘出对应人脸和汽车相关部件的视觉词组.

为了使视觉词组更加实用, Zhang 等 [36] 从谷歌图像搜索引擎中搜集了 376,500 幅关于 1,506 个物体或场景类别的图像, 并利用图像的类别信息来选取对类别具有更好区分力的视觉单词 (descriptive visual word, DVW) 和视觉词组 (descriptive visual phrase, DVP), 并在图像检索、物体识别和图像搜索重排序等问题中展现出明显的改进效果 (图例见文献 [36] 中图 8). 这个工作中的视觉词组主要由两个视觉单词组成, 要求在图像中相距较近且在图像库中出现频次较高, 这样相当于将视觉单词间的位置关系隐式地考虑进来. 为了进一步利用视觉单词间的空间位置, Zhang 等 [37] 将视觉单词在视觉词组中的位置显示地用于视觉词组间的相似度度量中, 对于提高图像检索的准确率有明显的效果.

由于这类方法大多基于 SIFT 特征, 通过视觉词组中引入位置关系更进一步增强了匹配的准确度, 因此这类方法对于检测重复图像和刚体类物体识别非常有效, 但是对于物体类别识别尤其是可变形物体的识别效果较为有限. 另一方面, 通过将空间位置信息编码到视觉词组中, 可以继续利用倒排索引的方法, 对于提高检索系统的效率非常有利.

4.4 基于局部结构的视觉模式

基于局部关键点的视觉词组多以两到三个视觉单词的组合为主, 并且要求词组之间的匹配为精确匹配, 这样有利于构建倒排索引. 相比之下, 基于局部结构的图像模式一般是指在一个局部区域中更多关键点的组合或通过其他特征来描述此局部区域的结构, 并且两个基于局部结构的视觉模式之间的匹配允许近似匹配.

例如, Wu 等在文献 [38] 中提出特征束 (bundling feature) 的方法，采用 MSER 的方法 [39] 检测图像中具有较强的仿射变换不变性的区域, 并将每个 MSER 区域所包含的 SIFT 特征点构成一个集合 (特征束) 来对该区域进行描述, 不同区域间的相似度用特征束间的重叠度来度量, 并考虑特征束中特征点的位置关系. 这种方法可以对图像中比较重要 (或最具有不变性) 的局部区域进行描述, 在大规模图像库中检测重复图像具有很高的准确度. 其不足之处在于特征束间的相似度较为复杂, 这种特征定义的空间不是度量空间, 导致索引和排序都较为复杂. 为了解决这个问题, Xu 等 [40] 提出 Nested SIFT 的方法, 利用 SIFT 特征点的尺度特性来挖掘中等尺度的特征点, 并利用其覆盖的小尺度特征点集来表征一个描述单元. 两组 Nested SIFT 间的相似度度量首先考虑中尺度特征点间的相似度, 以过滤不必要的小尺度特征点集间的相似度计算, 因而速度更快. 为了获取更为紧凑的 Nested SIFT 的表示, 作者将 Nested SIFT 中的特征点集转换为 SimHash 代码 [41] 并得到更快的匹配速度. Nested SIFT 对于通过多个图像进行三维重建和重复图像检测的应用问题具有高效准确的效果.

上述两种方法主要还是对重复图像检测比较有效, 这也是基于 SIFT 特征的模式表示的主要特点. 为了从图像数据中挖掘出更具有推广能力的视觉模式, 需要放松对 SIFT 特征的匹配要求, 并引入空间位置信息且允许足够的形变. 为此, Liu 等在文献 [42] 中展示了一种更灵活的从单幅图像中挖掘重复模式的方法, 为了描述某一重复模式在多个物体上出现的规律, 作者定义了一个特征点赋值矩阵, 一列对应一个物体, 一行表示一个视觉单词在每个物体上是否出现; 并且任意给定一个矩阵, 都可以评价该矩阵刻画的视觉模式是否合理. 在此基础上, 作者采用一种贪婪算法 GRASP (greedy randomized adaptive search procedure) 对矩阵进行增删和修改的操作, 以获取对应于尽可能多实例的赋值矩阵, 以

表示一个视觉模式. 该方法允许同一视觉模式的不同实例可以缺失部分特征点, 并且允许相当大的几何形变 (图例见文献 [42] 中图 6).

另外, 借助其他特征也可以对局部结构进行更具有推广能力的模式表示. 比如在文献 [43] 中, Zhang 等采用 SIFT, LBP, 颜色直方图等三种特征共同描述图像局部区域块, 并将同一类别中的局部区域块聚成若干类, 并利用类别信息为聚类结果标注来生成 ObjectWords, 并进而构成 ObjectBook. ObjectWords 可以对新图像进行分析生成语义描述, 并用于基于语义的图像检索. 而在文献 [44] 中, Yang 等采用 HOG 特征 [22] 对来自多个图像的关于同一类物体同一部件的局部区域训练分类器, 以检测新图像中是否有该部件从而得到物体是否存在的语义信息.

从这些工作中我们可以看出, 基于局部结构的视觉模式通常具有更灵活的模式匹配方式, 因而也具有更好的推广能力. 但是这类工作多是在较小规模的数据集学习和挖掘的, 如何更加有效地利用大规模数据学习更丰富更具有区分力的视觉模式, 仍有很多挑战需要解决.

4.5 基于隐层表示的视觉模式

长期以来, 好的特征都是靠人的经验设计出来的, 经过实际验证被大家广为接受的特征往往需要经过长时间的研究和探索. 以 SIFT 特征为例, 从研究人员开始关注立体视觉中图像匹配问题开始研究图像中角点的检测 [45], 到局部关键点尺度选择 [46], 再到工程上高效的实现方法 [12], 前后经过了近二十年的时间.

因此, 研究人员非常希望能够通过机器学习的方法, 从大量的数据中自动学习出隐层模式规律, 并用于有效的特征表示. 这方面的工作直到 2009 年才开始看到效果.

2009 年, Lee 等在文献 [47] 中提出采用卷积深度置信网络 (convolutional deep belief networks) 并通过无监督学习的方法, 从大量的图像数据中学习层级化的视觉特征表示. 这个方法的关键是采用了多层的神经网络, 并通过卷积层和池化 (pooling) 操作的方法学习图像中具有一定的平移不变性的局部特征. 通过三层的卷积和池化操作, 该神经网络的第一层、第二层及第三层可以分别学习出表示边缘检测、物体部件及整个物体的隐式模式 (图例见文献 [47] 中图 3), 这些模式刻画了大量图像数据中的局部性的隐式结构化信息, 对于从不同的粒度表示图像的语义有很大的启发和帮助.

2012 年, 多层神经网络在图像分类和模式提取方面取得更加令人瞩目的结果. 在文献 [48] 中, Le 等构造了一个 9 层局部连接神经网络, 对 1 千万幅网络图像进行无监督学习, 在顶层的众多神经元中发现了和人脸高度相关的神经元, 更多的实验结果显示其他的神经元也和一些视觉特征显著的概念相关. 这个工作有效地展示了从大规模无标注数据中学习隐含模式的可能性. 因为该网络神经元数量 (10^5) 和神经元之间的连接数量 (10^9) 极为庞大. 训练过程也采用了分布式计算的方法, 利用 1000 台服务器共计 16000 个 CPU 核完成计算.

同样在 2012 年, Krzhevsky 等 [49] 采用双 GPU 并行计算的方法, 构造了一个多层卷积神经网络, 在 ImageNet 1000 类分类问题的竞赛中取得最佳结果, 在 2012 年的数据集上其 5 选分类正确率达到 84.7% (而第二名为 73.8%). 值得注意的是, 这个工作采用了监督学习方法, 其顶层神经元直接对应于图像类别, 这些神经元对于图像的响应构成了一个区分度很好的语义空间, 在此空间中的图像检索效果更加接近语义上的相似度. 在这些神经网络中, 由于采用了局部连接的神经元和池化 (pooling) 层, 因而高层的神经元具有一定的平移不变性, 这对于模式匹配也是非常关键的特性.

近年来的这些研究工作, 表明了模式挖掘和机器学习之间结合的必要性和重要性. 由于图像数据的构成元素非常复杂, 只有从大规模的数据中才可以学习出规律性的视觉模式, 数据规模的增大和神

经网络复杂度的增高, 使得大规模高效计算成为这个研究方向的关键技术手段. 相信特征学习将会成为今后重要的研究方向, 对于寻找有效的中层表示以解决语义鸿沟问题有着重要的意义.

5 总结

在互联网时代, 爆炸式增长的数字图像不仅给图像检索带来了巨大的挑战, 也给图像内容分析带来了用之不尽的数据和不同以往的研究方法. 本文简单回顾了图像检索的三个阶段的研究历史, 以及在此过程中数据量的增多对图像检索带来的影响, 并对作为关键问题的特征提取方面的研究进行了深入的分析, 总结和介绍了近年来关于特征提取和表示方面的重要研究方向, 即寻找中层表示的视觉特征和视觉模式. 根据视觉模式的表征粒度, 本文将视觉模式分为五种不同的类别分别进行了介绍, 并强调大数据和图像关联文本对于挖掘具有一定的语义信息的视觉模式的重要性, 以及近年来通过机器学习的方法自动学习特征表示的显著进展.

由于图像检索具有高度的应用价值, 因此该问题将会长期受到工业界的关注, 尤其是在图像语义分析、高效索引、结果排序、用户点击数据的利用方面可以预期会有持续的进展. 另一方面, 作为基础研究问题, 如何从图像中提取更好的视觉特征、缩小和语义概念之间的差距, 将会是学术界长期关注的研究问题, 尤其是采用机器学习的方法自动学习视觉特征将会是近期的热点研究方向.

然而我们也应该看到, 尽管多层神经网络在大规模模式挖掘和计算机视觉领域已初获进展, 但这些研究工作仍处于起步阶段, 关于特征表示和学习的理论和方法也亟待更深入的研究. 特别值得深入的问题是, 如何从大数据中学习结构化的模式, 并使得这些模式具备一定的重复性 (在多个数据样本中出现)、不变性 (对于平移、尺度、旋转变换的鲁棒性) 和完备性 (可以构成尽可能完全的语义空间). 在大数据环境下, 对这些问题的研究还需要考虑噪声的影响和计算效率的问题, 这些问题都给机器学习的研究和图像内容分析带来了巨大的挑战.

参考文献

1 Zhang L, Rui Y. Image search from thousands to billions in 20 years (to appear). ACM Trans Multimedia Comput Commun Appl, 2013

2 Jain R. NSF workshop on visual information management systems. SIGMOD Record, 1993, 22: 57–75

3 Rui Y, Huang T, Ortega M, et al. Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Trans Circuits Syst Video Technol, 1998, 8: 644–655

4 Wang J. Encyclopedia of Data Warehousing and Mining. 2nd Ed. Hershey: IGI Global, 2009. 758–763

5 Wang J, Hua X. Interactive image search by color map. ACM Trans Intell Syst Technol, 2011, 3: 12

6 Cao Y, Wang H, Wang C, et al. Mindfinder: interactive sketch-based image search on millions of images. In: Proceedings of ACM International Conference on Multimedia, Florence, 2010. 1605–1608

7 Zha Z, Yang L, Mei T, et al. Visual query suggestion. In: Proceedings of ACM International Conference on Multimedia, Beijing, 2009. 15–24

8 Niblack C, Barber R, Equitz W, et al. The QBIC project: querying images by content using color, texture, and shape. In: Proceedings of IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology, San Jose, 1993. 173–187

9 Ma W Y, Manjunath B. NETRA: a toolbox for navigating large image databases. In: Proceedings of International Conference on Image Processing, San Antonio, 1997. 568–571

10 Smith J R, Chang S F. VisualSEEk: a fully automated content-based image query system. In: Proceedings of ACM International Conference on Multimedia, Seattle, 1997. 87–98

11 Smith J R, Chang S F. Visually searching the Web for content. IEEE Multimedia, 1997, 4: 12–20

12 Lowe D. Object recognition from local scale-invariant features. In: Proceedings of IEEE International Conference on Computer Vision, Kerkyra, 1999. 1150–1157
13 Gionis A, Indyk P, Motwani R. Similarity search in high dimensions via hashing. In: Proceedings of the International Conference on Very Large Data Bases, Edinburgh, 1999. 518–529
14 Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20: 273–297
15 Wang J, Li J, Wiederhold G. SIMPLIcity: semantics-sensitive integrated matching for picture libraries. IEEE Trans Pattern Anal Machine Intell, 2001, 23: 947–963
16 Quack T, Mönich U, Thiele L, et al. Cortina: a system for large-scale, content-based web image retrieval. In: Proceedings of ACM International Conference on Multimedia, New York, 2004. 508–511
17 Wang X J, Zhang L, Jing F, et al. AnnoSearch: image auto-annotation by search. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, New York, 2006. 1483–1490
18 Li X, Chen L, Zhang L, et al. Image annotation by large-scale content-based image retrieval. In: Proceedings of ACM International Conference on Multimedia, Santa Barbara, 2006. 607–610
19 Torralba A, Fergus R, Freeman W. 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans Pattern Anal Machine Intell, 2008, 30: 1958–1970
20 Wang X J, Zhang L, Liu M, et al. Arista-image search to annotation on billions of web photos. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 2010. 2987–2994
21 Rui Y, Huang T, Chang S F. Image retrieval: current techniques, promising directions, and open issues. J Visual Commun Image Representation, 1999, 10: 39–62
22 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Diego, 2005. 886–893
23 Shrivastava A, Malisiewicz T, Gupta A, et al. Data-driven visual similarity for cross-domain image matching. ACM Trans Graphics, 2011, 30: 154:1–154:10
24 Ahonen T, Hadid A, Pietikäinen M. Face recognition with local binary patterns. In: Proceedings of European Conference on Computer Vision, Prague, 2004. 469–481
25 Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Machine Intell, 2002, 24: 971–987
26 Wang X J, Xu Z, Zhang L, et al. Towards indexing representative images on the web. In: Proceedings of ACM International Conference on Multimedia, Naran, 2012. 1229–1238
27 Jaimes A, Chang S F, Loui A C. Duplicate detection in consumer photography and news video. In: Proceedings of ACM International Conference on Multimedia, Juan les Pins, 2002. 423–424
28 Chum O, Matas J. Large-scale discovery of spatially related images. IEEE Trans Pattern Anal Machine Intell, 2010, 32: 371–377
29 Wang B, Li Z, Li M, et al. Large-scale duplicate detection for web image search. In: Proceedings of IEEE International Conference on Multimedia and Expo, Toronto, 2006. 353–356
30 Lorenzo T, Szummer M, Fitzgibbon A. Efficient object category recognition using classemes. In: Proceedings of European Conference on Computer Vision, Heraklion, 2010. 776–789
31 Tsai D, Jing Y, Liu Y, et al. Large-scale image annotation using visual synset. In: Proceedings of IEEE International Conference on Computer Vision, Barcelona, 2011. 611–618
32 Bergamo A, Lorenzo T, Fitzgibbon A. Picodes: Learning a compact code for novel-category recognition. In: Proceedings of Neural Information Processing Systems, Granada, 2011. 2088–2096
33 Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos. In: Proceedings of IEEE International Conference on Computer Vision, Nice, 2003. 1470–1477
34 Yuan J, Wu Y, Yang M. Discovery of collocation patterns: from visual words to visual phrases. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007. 1–8
35 Han J, Cheng H, Xin D, et al. Frequent pattern mining: current status and future directions. Data Mining Knowl Discovery, 2007, 15: 55–86
36 Zhang S, Tian Q, Hua G, et al. Descriptive visual words and visual phrases for image applications. In: Proceedings of ACM International Conference on Multimedia, Beijing, 2009. 75–84

37 Zhang S, Huang Q, Hua G, et al. Building contextual visual vocabulary for large-scale image applications. In: Proceedings of ACM International Conference on Multimedia, Florence, 2010. 501–510
38 Wu Z, Ke Q, Isard M, et al. Bundling features for large scale partial-duplicate web image search. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009. 25–32
39 Jiri M, Chum O, Urban M, et al. Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of British Machine Vision Conference, Cardiff, 2002. 384–393
40 Xu P, Zhang L, Yang K, et al. Nested-SIFT for efficient image matching and retrieval. IEEE Multimedia, 2013, 20: 34–46
41 Charikar M. Similarity estimation techniques from rounding algorithms. In: Proceedings of ACM Symposium on Theory of Computing, Montreal, 2002. 380–388
42 Liu J, Liu Y. GRASP recurring patterns from a single view (to appear). In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Oregon, 2013
43 Zhang S, Tian Q, Huang Q, et al. Objectbook construction for large-scale semantic-aware image retrieval. In: Proceeding of IEEE International Workshop on Multimedia Signal Processing, Hangzhou, 2011. 1–6
44 Yang K, Zhang L, Rui Y, et al. PartBook for image parsing. In: Proceedings of IEEE Computer Society Workshop on Perceptual Organization in Computer Vision, Providence, 2012. 17–24
45 Moravec H. Rover visual obstacle avoidance. In: Proceedings of International Joint Conference on Artificial Intelligence, Vancouver, 1981. 785–790
46 Lindeberg T. Feature detection with automatic scale selection. Int J Comput Vision, 1998, 30: 79–116
47 Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of ACM International Conference on Machine Learning, Montreal, 2009. 609–616
48 Le Q, Ranzato M, Monga R, et al. Building high-level features using large scale unsupervised learning. In: Proceedings of ACM International Conference on Machine Learning, Edinburgh, 2012. 81–88
49 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of Neural Information Processing Systems, Lake Tahoe, 2012. 1097–1106

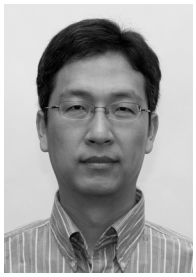
Large-scale web image search and pattern mining

ZHANG Lei

Microsoft Search Technology Center Asia, Beijing 100080*, China*
E-mail: leizhang@microsoft.com

Abstract The explosive growth of web images not only brings many technical challenges to image search, but also provides almost unlimited training data and new ideas to various computer vision problems. This paper presents a brief historical review of three stages of image retrieval, with a particular emphasis on the impact of large-scale web images to image retrieval. Based on the review, the paper discusses the fundamental problem of feature extraction in image retrieval, and the recent research trend on visual pattern mining to bridge the semantic gap. According to their representation granularity, the paper divides visual patterns into five categories and introduces their related work respectively, which also shows the great importance of big data to visual pattern mining.

Keywords image retrieval, visual pattern, pattern mining, content analysis, big data



ZHANG Lei earned his B.E. and M.E. in computer science from Tsinghua University in 1993 and 1995. After two years working in industry, he later returned to Tsinghua and received his Ph.D. degree in computer science in 2001. Before 2013, he was a senior researcher in the Multimedia Search & Mining Group at Microsoft Research Asia. He is interested in research problems of image search, Internet vision and information retrieval, and holds 20 U.S. patents for his innovation in these fields. He is now a senior research manager in Microsoft Search Technology Center Asia. He is an IEEE senior member and an ACM senior member, and has served as an associate editor of Multimedia System Journal, program area chair of ACM Multimedia 2012, 2013, ICPR 2012 and ICME 2011, and also served on international conference program committees, including ACM Multimedia, ICCV, CVPR, WWW, SIGIR, etc.