计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20180058
55(9):1829-1842,2018

基于深度特征的无监督图像检索研究综述

张 皓 吴建鑫
(计算机软件新技术国家重点实验室(南京大学) 南京 210023)
(wujx2001@nju.edu.cn)

A Survey on Unsupervised Image Retrieval Using Deep Features

Zhang Hao and Wu Jianxin
(*National Key Laboratory for Novel Software Technology* (*Nanjing University*), *Nanjing* 210023)

Abstract Content-based image retrieval (CBIR) is a challenging task in computer vision. Its goal is to find images among the database images which contain the same instance as the query image. A typical image retrieval approach contains two steps: extract a proper representation vector from each raw image, and then retrieve via nearest neighbor search on those representations. The quality of the image representation vector extracted from raw image is the key factor to determine the overall performance of an image retrieval approach. Image retrieval have witnessed two developing stages, namely hand-craft feature based approaches and deep feature based approaches. Furthermore, there are two phases in each stage, i.e., one phase of using global feature and another phase of using local feature based approaches. Due to the limited representation power of hand-craft features, nowadays, the research focus of image retrieval has shifted to how to make the full utility of deep features. In this study, we give a brief review of the development progress of unsupervised image retrieval based on different ways to extract image representations. Several representative unsupervised image retrieval approaches are then introduced and compared on benchmark image retrieval datasets. At last, we discuss a few future research perspectives.

Key words image retrieval; deep learning; convolutional neural networks; computer vision; unsupervised learning

摘 要 基于内容的图像检索(content-based image retrieval, CBIR)是一项极具挑战的计算机视觉任务.其目标是从数据库图像中找到和查询图像包含相同实例的图像.一个典型的图像检索流程包括2步:设法从图像中提取一个合适的图像的表示向量和对这些表示向量进行最近邻搜索以找到相似的图像.其中,决定图像检索算法性能的关键在于其提取的图像表示的好坏.图像检索中使用的图像表示经历了基于手工特征和基于深度特征两大时期,每个时期又有全局特征和局部特征2个阶段.由于手工特征的表示能力有限,近年来图像检索的研究主要集中在如何利用深度特征.将以提取图像表示的不同思路为线索,回顾无监督图像检索领域的发展历程,介绍该领域的一些代表性算法,并比较这些算法在常用数据集上的性能表现,最后探讨未来的研究方向.

关键词 图像检索;深度学习;卷积神经网络;计算机视觉;无监督学习

中图法分类号 TP183

收稿日期:2018-01-24;**修回日期**:2018-06-08

基金项目:国家自然科学基金优秀青年科学基金项目(61422203)

This work was supported by the National Natural Science Foundation of China for Excellent Young Scientists (61422203).

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

基于内容的图像检索(content-based image retrieval, CBIR)是一项极具挑战的计算机任务,并且得到了长期的研究关注[1-4].给定一个包含特定实例(例如特定目标、场景、建筑等)的查询图像,图像检索旨在从数据库图像中找到包含相同实例的图像[5].但由于不同图像的拍摄视角、光照或遮挡情况不同,如何设计出能应对这些类内差异的有效且高效的图像检索算法仍是一项研究难题.

一个典型的图像检索流程包括2步:设法从图像中提取一个合适的图像的表示向量,和对这些表示向量用欧氏距离或余弦距离进行最近邻搜索以找到相似的图像.可以看出,决定一个图像检索算法性能的关键在于提取的图像表示的好坏.

对图像检索的研究至今已有约20年的时间.图像检索中使用的图像表示经历了基于手工特征和基于深度特征两大时期,每个时期又有全局特征和局部特征2个阶段.由于手工特征的表示能力有限,近年来图像检索的研究主要集中在如何利用深度特征.一些早期工作直接提取深度全连接特征作为图像的表示向量,这实质是对图像整体语义信息进行描述的深度全局特征.然而,由于全局特征缺乏对图像细节的描述,后来研究关注点集中到深度局部特征.基于深度局部特征的图像检索可以分为3类:基于局部表示聚合、基于深度卷积特征聚合和基于多层融合的方法.本文将以提取图像表示的不同思路为线索,介绍无监督图像检索领域的一些代表性算法,并探讨未来可能的研究方向.

本文首先对无监督图像检索做一概述,介绍一些比较常用的数据集,对一些基于手工特征的早期算法作一简要回顾,并分别介绍一些基于深度全局特征和深度局部特征的代表算法.之后,我们给出一些能提升图像检索性能的实现细节,并比较各算法在图像检索常用数据集上的性能.最后,我们对图像检索领域未来可能的发展方向及其挑战进行展望.

1 无监督图像检索概述

图像检索旨在从数据库图像中找到包含相同实例的图像[5].其中图像表示的好坏对一个图像检索算法的性能起决定性影响.回顾整个发展历程,图像检索中使用的图像表示经历了基于手工特征和基于深度特征两大时期,每个时期又有全局特征和局部特征2个阶段,如图1所示.最初图像检索主要基于一些手工全局图像特征,如颜色[6]、边缘[6]、纹理[7]、GIST[8-9]等.但由于这些全局图像特征容易受图像中的光照条件、位移、遮挡、截断等因素影响,后来研究焦点逐渐转向以SIFT[10]特征结合BoW聚合[11]为代表的基于手工局部特征的图像表示提取方法.然而,由于特征的表示能力有限,检索性能也往往面临很大的局限性.

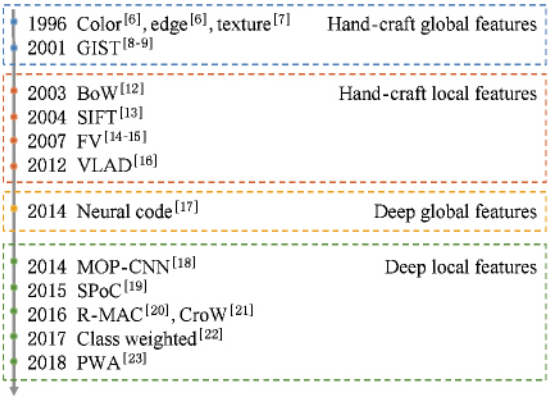

Fig. 1 Four developing stages of unsupervised image retrieval and representative approaches

图1 无监督图像检索的4个发展阶段及其代表方法

深度学习利用多层非线性变换的堆叠来对数据的高层次表示进行建模.由于从大规模数据集如ImageNet[24]上预训练好的深度卷积神经网络中可以提取通用的图像表示,并用于其他视觉任务,包括图像检索[25-27].因此,近年来图像检索的研究主要集中在如何利用深度卷积神经网络的图像表示.

由于图像检索经常需要面对大量动态变化或流式数据库图像,对这些图像进行标注代价十分昂贵.因此,从零训练或者微调一个预训练网络既不高效也不具备可扩展性.本文旨在对不借助其他监督信息,只利用ImageNet预训练模型作为固定的特征提取器来提取图像表示的研究方法进行综述.

由于深度全连接特征提供了对图像内容高层级的描述且是"天然"的向量形式,一些早期工作直接提取深度全连接特征作为图像的表示向量.这种深度全连接特征实质是对图像整体语义信息进行描述的深度全局特征.但由于全局特征旨在进行图像分类,缺乏对图像细节的描述,后来,研究关注集中到深度局部特征.

基于深度局部特征的图像检索研究可大致分为3类:局部表示聚合、深度卷积特征聚合和多层融合.基于局部表示聚合的方法先设法从输入图像中提取一系列的局部区域,之后分别将这些局部区域前馈网络生成对应的局部表示,最后通过特定编码

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

方法将这些局部表示聚合为最终表示.这类方法的弊端在于需要前馈网络多次.深度卷积特征可以看作是对图像局部区域的描述,基于深度卷积特征聚合的方法只前馈网络一次生成深度卷积特征,并对其聚合为图像最终表示.此外,由于不同层的特征包含不同层级的语义信息,基于多层融合的方法旨在使不同层特征互补.

2　常用数据集

本节简要介绍图像检索领域若干常用数据集及评价指标.

Oxford5k 数据集[28]是从 Flickr 获得的哈佛大学 11 个地标建筑的 5 063 张图像,每个地标建筑包含 5 张查询图像,即共有 55 张查询图像.每个查询图像有一个手工标注的边界框.对每个查询图像,数据库图像被分成了 4 类:good,ok,junk 和 bad.前两者被认为是匹配的图像,而后两者被认为是无关图像.这些建筑风格十分接近,所以 Oxford5k 是一个比较具有挑战性的数据集.

Paris6k 数据集[29]是从 Flickr 获得的巴黎 11 个地标建筑的 6 412 张图像,每个地标建筑包含 5 张查询图像,即共有 55 张查询图像.和 Oxford5k 数据集标注方法相同,每个查询图像有一个手工标注的边界框,且对每个查询图像,数据库图像被分成了 4 类:good,ok,junk 和 bad.该数据集建筑物风格比 Oxford5k 更多样.

此外,从 Flickr 中爬取的属于 145 个常见类别的 99 782 张无关图像可以加入 Oxford5k 和 Paris6k,这样分别构成了 Oxford105k 和 Paris106k 数据集.这 2 个数据集图像规模大、种类多,通常用于检测算法在存在大量无关图像情况下的检索效果.

Holidays 数据集[30]是从个人相册中收集的 1 491 张风景图,并根据内容场景分为 500 组.每组包含 1 张查询图像,即共有 500 张查询图像.由于其中有些图像并不是自然的朝向,即被旋转了 90°,许多方法会手工将这些图像旋转为正常的朝向[17].这通常会带来 2%～3%的性能提升.该数据集包含了不同风景、场景和遗迹等,多样性较 Oxford5k, Paris6k 数据集更大.

目前在图像检索中最常用的评价指标是平均准确率均值(mean average precision, mAP),其计算方法如下:对每张查询图像,我们根据查询图像和数据库图像的表示向量间的距离,可以对数据库图像产生一个排序.进而,我们可以画出对该查询图像的查准率-查全率(P-R)曲线,平均准确率(average precision, AP)对应于 P-R 曲线下的面积.对所有查询图像的 AP 做平均,即可得到 mAP,取值为 0%～100%.

此外,Ukbench 数据集[13]有时也会被使用.Ukbench 包含了 10 200 张室内照片,并根据内容分成 2 550 组.和上文介绍的数据集不同,Ukbench 没有专门留出查询图像,实践中通常是让 10 200 张图像轮流作为查询图像.并且,评价指标一般不用 mAP,而是计算对每个查询图像最接近的 4 张数据库图像中有几张属于相同类别,取值为 0～4,即前 4 检索结果的查准率乘以 4.

3　基于手工特征的早期算法简述

对图像检索的研究至今已有约 20 年的时间.从 20 世纪 90 年代到本世纪初,图像检索主要基于一些手工全局图像特征,如颜色[6]、边缘[6]、纹理[7]、GIST[8-9]等.然而,这些全局图像特征容易受图像中的光照条件、位移、遮挡、截断等因素影响.

图像检索是最早采纳 BoW 编码的领域[21].在 2003 年 BoW 聚合得到计算机视觉领域的广泛关注[11],而 SIFT 特征在 2004 年被发明[10].随即,研究焦点逐渐由直接使用手工全局特征转向以 SIFT 特征结合 BoW 聚合为代表的基于手工局部特征的图像表示提取方法.

相比经典手工全局特征,SIFT 特征受图像旋转、尺度变换、光照条件改变等影响较小.此外,SURF[31]特征也比较常用.另一方面,在有了图像局部特征之后,我们需要某种聚合方法将这些局部特征汇总为图像表示.BoW 受文本处理的启发,其动机是相似的文本中应包含相似的词[32].BoW 将图像的局部特征作为一个个视觉词,其忽略各视觉词之间的关系,将一幅图像表示为包含了若干视觉词的“袋子”.“袋子”中包含的视觉词可以通过词典映射构成图像的表示向量,其中每维记录了该“袋子”中特定视觉词的出现频数.

除 BoW 聚合外,后来产生了 VLAD[16],FV[14-15]和 Triangulation embedding[33]聚合技术,并取得了比经典 BoW 聚合更好的性能.VLAD 先对局部图像特性进行 k-均值聚类,之后将每个局部特征根据其与最近聚类中心的距离进行编码.FV 和 VLAD

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

类似,但 FV 使用高斯混合模型[34]将局部图像特性对应到视觉词中. 此外,FV 还使用了二阶信息. Triangulation embedding 同样先进行聚类,而之后的编码是依据每个局部特征与所有聚类中心的归一化距离.

4 基于深度全局特征的图像检索研究

深度卷积神经网络通过多层非线性变换的嵌套来对图像的高层级特征进行建模. 深度卷积神经网络在 2012 年 ILSVRC 竞赛取得巨大突破后[35],得到了计算机视觉领域的广泛关注. 从大规模数据集如 ImageNet[18]上预训练好的深度卷积神经网络中可以提取通用的图像表示,并用于其他视觉任务,包括图像检索[25-27]. 因此,近年来,虽然基于手工特征的方法仍在发展[36-37],图像检索的研究焦点逐渐由经典手工特征过渡到利用深度卷积神经网络的中间层特征,并取得了比经典手工特征更优异的性能.

深度神经网络的全连接层特征提供了对图像内容高层级的描述. 此外,由于全连接层特征是"天然"的向量形式,一些早期工作直接将整张图像输入预训练好的网络,并提取深度全连接特征作为图像的表示向量,如图 2 所示. 这种深度全连接特征实质是对图像整体语义信息进行描述的深度全局特征.

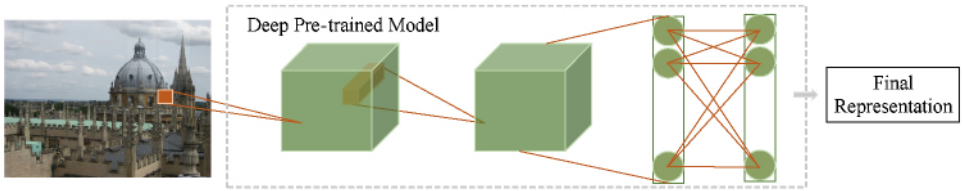

Fig. 2 Pipeline of image retrieval using deep global features

图 2 使用深度全局特征进行图像检索的流程图

文献[17]发现即使用于预训练的数据集(如 ImageNet)和用于图像检索的数据集有很大差异,由于深度卷积神经网络学得的特征有很强的泛化能力,用 ImageNet 预训练模型提取的深度全连接特征仍能取得很好的图像检索效果. 此外,为了得到更精简的图像表示,我们可以用 PCA 对从深度全连接特征提取的图像表示进行降维. 即使图像表示从 4 096 维降到 128 维,仍然能取得超过经典手工特征的检索性能[17].

另一方面,由于预训练数据集和用于图像检索的数据集差异的存在,文献[17,38]发现使用 fc7 特征会比 fc8 特征有更好的检索效果. 这是因为非常高层的深度特征旨在进行预训练数据集的分类任务,而略低层的深度特征有更好的对其他数据集的泛化能力.

使用深度全局特征作为图像表示十分简单直接,然而,全局特征旨在进行图像分类,缺乏对图像细节的描述. 此外,深度全局特征对图像平移、旋转和尺度缩放比较敏感[18]. 综上,使用全局特征并不是一个理想的选择. 后来,研究关注集中到深度局部特征.

5 基于深度局部特征的图像检索研究

本节将对有代表性的基于深度局部特征的图像检索工作做一简要回顾. 根据图像表示提取流程的不同,这些方法可以分为 3 类:局部表示聚合、深度卷积特征聚合和多层融合.

5.1 局部表示聚合

受经典 SIFT 特征和 BoW 聚合思路的启发,一些工作用深度特征代替经典 SIFT 特征,并对经典编码技术如 BoW, VLAD, FV 等加以改进. 这类方法先设法从输入图像中提取一系列的局部区域,之后分别将这些图像局部区域前馈网络,并生成对应的局部图像表示. 最后,通过特定聚合方法将这些局部图像表示聚合为最终图像表示,如图 3 所示.

这类方法的关键是如何从输入图像中提取局部区域,以及如何对局部表示进行聚合. 根据提取局部区域方法不同,我们进一步将这类方法分为 3 类:基于滑动窗的局部区域提取、兴趣区域检测和基于候选区域(region proposal)的局部区域提取.

1) 基于滑动窗的局部区域提取. 这类方法使用某一特定的滑动窗在输入图像上滑动,并提取滑动窗口在输入图像不同位置的对应图像局部区域. 此外,由于图像中的目标可能会有多种大小,这类方法通常会用一系列不同大小的滑动窗分别在输入图像上进行滑动以生成局部区域.

文献[39]使用了 4 种不同大小的滑动窗并使用 fc7 特征作为图像的局部表示. 并且文献[30]没有将

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

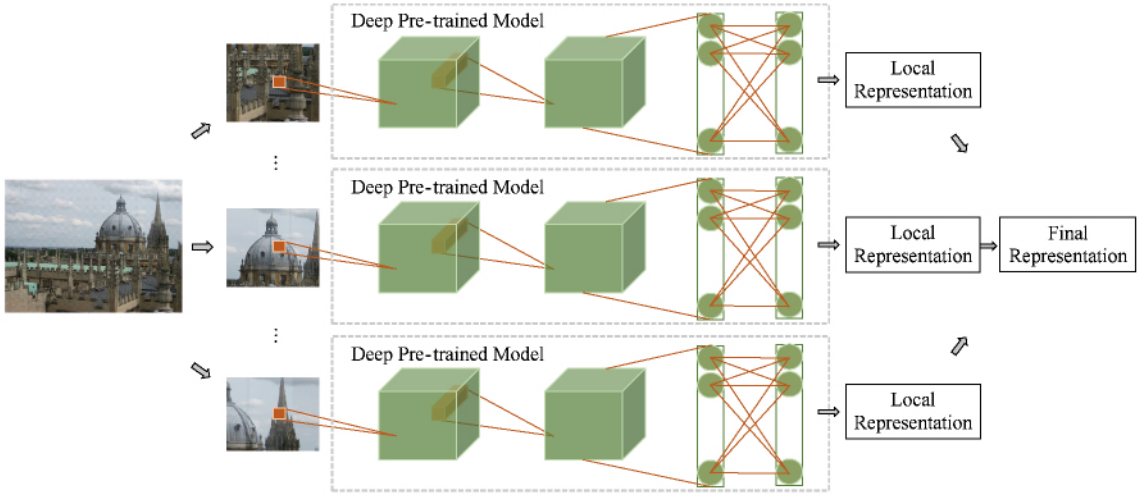

Fig. 3　Pipeline of image retrieval using deep local representation aggregation

图 3　使用深度局部表示聚合进行图像检索的流程图

局部表示聚合为全局表示,而直接用局部表示来进行图像匹配.对查询图像的某一局部区域和一张数据库图像,文献[39]计算该查询图像局部表示向量对数据库图像所有局部表示向量的欧氏距离,并定义其最小值作为查询图像局部区域到数据库图像的距离.进而,文献[39]定义查询图像所有局部区域到数据库图像距离的平均值作为查询图像到数据库图像的距离.然而,与将局部表示聚合为全局表示并利用全局表示进行相似度度量相比,基于局部表示的图像匹配需要较大的计算和存储开销.

MOP-CNN[18]使用了 3 种大小的滑动窗.此外,MOP-CNN 把对应 3 种滑动窗的 fc7 局部表示分别进行 VLAD 聚合,并级联(concatenate)得到最终的图像表示.MOP-CNN 可以看作是 SPM[40]利用深度特征的改进.SPM 综合了从全局到局部不同尺度图像区域信息,而 MOP-CNN 通过滑动窗提取 3 种尺度图像局部表示,其中最大尺度是从原图提取深度全局特征.MOP-CNN 同时利用了深度局部特征和深度全局特征,并取得比只用深度全局特征更好的性能.

2) 兴趣区域检测.这类方法利用某种特定的兴趣区域检测算法提取图像中的兴趣区域,再将兴趣区域逐一前馈网络生成图像的局部表示.兴趣点检测器能够检测出图像中对特定变换不敏感的图像区域[41].这使得即使在不同视角或光照条件下,对相同目标拍摄的 2 张图像有相似的兴趣区域.

Patch-CKN[42]使用 Hessian-affine 检测器[43]提取图像中的兴趣区域.Hessian-affine 检测器可以在图像中提取对仿射变换不敏感的兴趣区域.之后,兴趣区域中每个点的近邻通过仿射和朝向规范化构成网络的输入.最后,Patch-CKN 使用深度卷积核网络[44]生成深度局部表示,并使用 VLAD 聚合得到图像表示.

3) 基于候选区域的局部区域提取.受目标检测启发,这类方法利用某种特定的无监督候选区域生成算法得到可能包含目标的局部候选区域,再将这些局部区域前馈经过网络提取图像局部表示.

文献[45]使用了 selective search[46]提取图像 2 000 个局部区域并前馈网络,并对生成的 fc7 特征每维进行最大汇合(max-pooling)[47]得到图像表示.文献[45]发现,最少只使用 100 个候选区域就可以达到相当有竞争力的效果.

CCS[48]使用 EdgeBox[49]提取图像 100 个局部区域并前馈网络,再用 VLAD 对深度卷积特征进行聚合.此外,CCS 同时结合了经典 SIFT 特征和深度特征,以达到不同尺度(场景级别、目标级别、和点级别)特征互补的目的.

纵观这 3 类方法,为了精确找到有价值的局部区域,基于滑动窗的方法需要用一系列不同大小的滑动窗口,这需要消耗不小的计算开销.而基于兴趣区域检测和基于候选区域的方法由于只需提取一部分局部区域前馈网络,因此效率相比较高.基于兴趣区域检测的方法沿用了经典图像检索/计算机视觉的特征提取思路,而基于候选区域的方法受到目标检测任务做法的启发.但这些方法都需要多次前馈网络.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

5.2 深度卷积特征聚合

基于局部表示聚合的方法通常需要对多个图像局部区域分别前馈网络以生成深度局部特征，所以算法效率成为其瓶颈. 另一方面，由于深度卷积特征可以被看作是对图像局部区域的描述，其局部区域大小为该深度卷积特征的感受野[50]. 因此，这类方法只前馈网络一次生成深度卷积特征，并对其进行聚合以得到图像的最终表示，如图 4 所示. 图 4 中灰色的部分代表这类方法对预训练模型没有用到的部分(即全连接层部分).

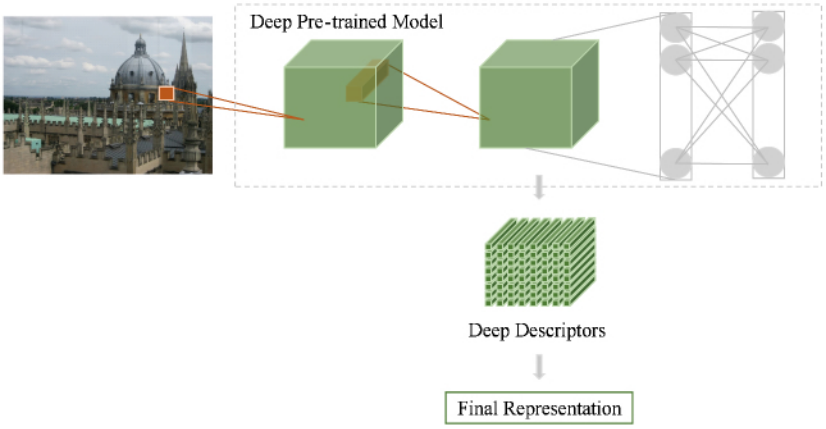

Fig. 4 Pipeline of image retrieval using aggregation of deep convolutional feature

图 4 使用深度卷积特征聚合进行图像检索的流程图

相比深度全局特征，深度卷积特征对图像的平移、裁剪、遮挡等更不敏感，并且保留了更多的图像细节信息[46]. 相比基于局部表示聚合的方法，基于深度卷积特征聚合的方法共享深度特征提取部分，只需前馈网络一次. 此外，使用深度卷积特征的另一个好处是可以处理任意大小的图像输入. 因此，这类方法逐渐成为近年来深度图像检索的主流方法.

这类方法的关键是如何对深度卷积特征进行聚合. 根据聚合时是否加权，我们可以将这类方法分为直接聚合和加权聚合 2 类. 下面我们对这 2 类方法分别予以综述.

本节使用如下符号. 给定输入图像 $\boldsymbol{I}$，深度卷积特征表示为 $\boldsymbol{A}\in \mathrm{R}^{D\times H\times W}$. 其可以理解为一系列二维的特征图 $\{\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_D\}$，也可以理解为包含 $H\times W$ 个网格区域，每个区域是一个 D 维的深度描述向量，记为 $\{\boldsymbol{x}_{11},\boldsymbol{x}_{12},\cdots,\boldsymbol{x}_{HW}\}$. 最终的图像表示向量记为 $\phi(\boldsymbol{I})$.

1) 直接聚合. 这类方法采用特定聚合方法对深度卷积特征进行聚合，以得到最终的图像表示. 聚合可以采用经典的 BoW, VLAD, FV 等，也可以使用(全局)最大汇合(max-pooling)或求和汇合(sum-pooling).

R-MAC[20] 对深度卷积特征进行滑动窗采样，并用最大汇合从这些采样得到的区域中提取图像的局部表示：

$$\phi(\boldsymbol{I})_{\Omega}=\max_{(i,j)\in\Omega}\boldsymbol{x}_{ij},$$

其中，Ω 代表局部卷积区域. 和 MOP-CNN 相比，R-MAC 也使用了 3 种不同大小的滑动窗. 不同之处在于，MOP-CNN 在输入图像上采样而 R-MAC 在深度卷积特征上采样，这使得 R-MAC 只需前馈网络一次提取深度特征，比 MOP-CNN 更加高效.

文献[50]使用空间最大汇合以提取固定大小的深度特征. 这样，即使输入图像的尺寸不一，提取得到的深度特征维数是一致的. 文献[50]采用了 1×1 和 2×2 两种尺度的空间最大汇合，并发现用再大的尺寸会损失性能. 文献[50]用 4 种图像尺寸的深度局部表示来进行图像匹配. 和文献[30]使用的距离度量相同，文献[50]定义查询图像局部区域到数据库图像的距离为最小查询图像局部表示向量和数据库图像局部表示的距离，及定义所有查询图像局部区域到数据库图像的距离的平均值作为查询图像到数据库图像的距离. 这适用于图像中目标可以任意大小、出现在任意位置的情形.

文献[52]提取深度卷积层的特征，并用 VLAD 进行编码得到图像的表示向量. 文献[52]发现随着提取的特征由浅层特征逐渐到深层特征，检索性能先上升后下降. 这是因为，越深层次的特征提取的语义信息越丰富，但是过深层次的特征包含的细节信息不足，对检索的判别能力下降. 并且，对不同数据集，最适合用于提取特征的卷积层也会不同. 总体来说，VGG-16 中 conv5-1 特征更适合用于图像检索.

BLCF[53] 提取深度卷积层的特征，并用 BoW 进

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

行编码得到图像的表示向量.使用 BoW 编码得到的表示向量更加稀疏,在实际中可以利用倒排索引,以达到比 VLAD 更快的检索速度.BLCF 使用 25 000个均值聚类中心.此外,BLCF 对深度卷积特征进行双线性差值上采样,以获得更大的深度卷积特征,并避免大尺寸图像输入带来的计算负担.但是,当数据集中有大量无关图像存在时,使用 BoW 编码的局限逐渐显露.

SPoC[19]发现深度卷积特征和 SIFT 特征虽然都是局部特征,但是有不同的性质.经典的编码方法如 VLAD,FV 和 Triangulation embedding 旨在提升经典局部特征如 SIFT 的判别能力.另一方面,由于深度卷积特征有比经典手工局部特征更高的判别能力,因此,适用于 SIFT 特征的编码方式不能简单移植到深度卷积特征.SPoC 发现,直接对深度卷积特征做求和汇合

$$\phi(\boldsymbol{I}) = \sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{x}_{ij}$$

简单有效,并可取得比其他常用聚合方法更好的性能.文献[51]发现,当图像中目标比较大时,使用求和汇合比最大汇合更好,而当目标比较小时,其背景噪声会干扰求和汇合的结果.

2) 加权聚合.这类方法延续了直接聚合对深度卷积特征进行编码的策略,并且在汇合时根据不同位置特征的重要性对深度卷积特征进行加权.我们使用 $\boldsymbol{\alpha}\in\mathrm{R}^{H\times W}$ 表示空间权重,$\boldsymbol{\beta}\in\mathrm{R}^{D}$ 表示通道权重,则图像表示为

$$\phi(\boldsymbol{I}) = \boldsymbol{\beta}\odot\sum_{i=1}^{H}\sum_{j=1}^{W}\alpha_{ij}\boldsymbol{x}_{ij},$$

其中,$\odot$ 代表矩阵的 Hadamard 乘法,即矩阵对应元素相乘.

SPoC[19]认为图像中的目标倾向于集中在图像的几何中心,因此 SPoC 在求和汇合时使用了一个高斯权重:

$$\alpha_{ij} = \exp\left(-\frac{\left(i-\frac{H}{2}\right)^2+\left(j-\frac{W}{2}\right)^2}{2\sigma^2}\right),$$

其中,σ 设定为图像中心到图像最短边界距离的 1/3.当特征接近图像中心时,其对应权重较高.和直接求和汇合相比,使用高斯权重能显著提升性能.SPoC 中使用的权重依赖于先验知识,和数据无关,而下述方法旨在从深度卷积特征中得到用于汇合的权重.

SCDA[54]利用深度卷积特征分布式表示的特性来得到空间权重.SIFT 特征结合 BoW 编码得到的图像表示向量每维对应一个明确的概念.与之不同的是,卷积特征中的神经元和概念之间是一个多对多的映射,即每个语义概念由不同神经元表示,而每个神经元又参与到许多不同概念的表示中去[55-56].SCDA 认为,虽然一个神经元的响应值对判断对应区域是否包含目标用处不大,但如果多个神经元同时有很大的响应值,那么该区域很有可能包含目标.

因此,SCDA 把特征图沿通道(channel)方向相加,得到一张 2 维的"聚合图":

$$\boldsymbol{S} = \sum_{d=1}^{D}\boldsymbol{F}_d,$$

之后,根据聚合图元素是否超过聚合图的均值

$$\mu = \sum_{i=1}^{H}\sum_{j=1}^{W}S_{ij}$$

将其 0/1 二值化,并将其作为求和汇合的权重

$$\alpha_{ij} = I(S_{ij}\geqslant\mu),$$

其中,$I(\cdot)$ 代表指示函数.SCDA 发现,即使预训练模型使用的数据集和检索的数据集差异很大,利用该空间权重,可以很好地分割出前景目标,并去除背景噪声的影响.

CroW[21]同时用了空间和通道方向的加权.空间权重用于凸显高度活跃的响应值,空间方向的权重是归一化并根号规范化后的聚合图:

$$\alpha_{ij} = \sqrt{\frac{S_{ij}}{\|\boldsymbol{S}\|_{\mathrm{F}}}}.$$

如果在某个空间位置有多个通道的神经元都有比较大的响应,则该位置包含目标的可能性更大,其对应空间权重 α_{ij} 也更大.在实验中发现,大的空间权重 α_{ij} 对应于图像中的显著视觉区域,

CroW 的通道权重用于应对"视觉爆发"[57]现象,即图像中多处出现几乎相同的局部特征,而这些特征将主导相似度度量.通道权重根据特征图的稀疏性定义,其类似于自然语言处理中 TF-IDF 特征中的 IDF 特征,用于提升不常出现但具有判别能力的特征.CroW 首先计算每个特征图非零元素个数:

$$Q_d = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}I(F_{dij}>0),$$

之后通道权重定义为

$$\beta_d = \ln\frac{\sum_{k=1}^{D}Q_k}{Q_d},$$

其中,为了表述简洁,略去了提升数值稳定性的项.

文献[58]发现,经典编码方法如 VLAD 和 FV 效果不如直接汇合的原因在于经典编码方法没有考虑空间权重.这些方法同等看待所有局部表示向量,这将会使背景噪声影响最终得到的图像表示.因此,文献[58]向经典 FV 编码引入空间加权,提出

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

SWFV 和 TSWFV 编码方法. SWFV 使用了 CroW 的空间权重 α 来凸显图像中的前景目标. TSWFV 使用了一个阈值以对低于该阈值的 CroW 空间权重置零,旨在完全去除背景噪声的干扰. TSWFV 先将空间权重由大至小排序 $(\alpha'_1,\alpha'_2,\cdots,\alpha'_{HW})$,之后根据预先指定的阈值 T 得到需要保留的元素个数:

$$N=\arg\min_n n,$$

$$\text{s.t.}\ \sum_{n=1}^{N}\alpha'_n>T,$$

并将排序超过 N 的权重置零.

PWA[23] 发现,深度卷积特征的不同通道对应于目标不同部分的响应. 因此,PWA 选取一系列有判别能力的特征图,将其归一化之后的结果作为权重分别进行汇合,并将其结果级联起来作为最终图像表示:

$$\alpha_{ij}=\sqrt{\frac{F_{dij}}{\|\boldsymbol{F}_d\|_{\mathrm{F}}}},$$

PWA 根据各特征图在数据库图像上的方差选择若干特征图来计算空间权重. PWA 首先将深度卷积特征进行求和汇合,之后分别计算各维在数据库图像上的方差,并选择方差最大的若干维对应的特征图计算空间权重. PWA 认为,这些方差最大的特征图在不同目标上有不同的相应,因此具有很强的判别能力.

上述方法的权重均与类别信息无关,而文献[22]试图结合网络的类别预测信息来使空间权重更具判别能力. 具体来说,文献[22]利用了 CAM[59] 来获取预训练网络中对应各类别的对具代表性区域的语义信息. 对于给定类别 k,其 CAM 结果是特征图对后续全连接层参数的线性组合:

$$\boldsymbol{C}_k=\sum_{d=1}^{D}w_{kd}\boldsymbol{F}_d,$$

其中,w_{kd} 是对应于第 k 个类、第 d 个特征图的全连接层参数. 而文献[22]使用归一化的 CAM 结果作为空间权重:

$$\alpha_{ij}=\sqrt{\frac{C_{dij}}{\|\boldsymbol{C}_d\|_{\mathrm{F}}}},$$

而通道权重采用和 CroW 相同的做法.

此外,根据选哪些类别进行 CAM 的方式不同,文献[22]进一步可分为 OnA 和 OfA 两种策略. 在 OnA 中,使用对查询图像预测概率最高的几个类别提取 CAM,进而对深度卷积特征进行加权. 然而,这会影响查询速度,并且 OnA 需要事先将数据库图像对应所有类别的 CAM 计算并保存到硬盘,这是很大的计算和存储开销. 另一方面,在 OfA 中,使用对数据库图像各自预测概率最高的几个类别提取 CAM. OfA 比 OnA 有更强的可扩展性,但 OfA 的检索性能不如 OnA.

5.3 多层融合

深度特征具有层次性,即从低层到高层是由纹理特征到高层语义特征的转变. 随着深度的增加,神经元的感受野随之扩大,更倾向于捕获全局的语义信息. 这类方法旨在使深度神经网络中不同层特征的信息互补,以综合不同层特征的不变性和判别能力. 由于不同层特征的感受野不同,多层特征融合可以同时捕获不同尺度下的图像语义信息[60-61],如图 5 所示. 图 5 中灰色的部分代表这类方法对预训练模型没有用到的部分(即全连接层部分).

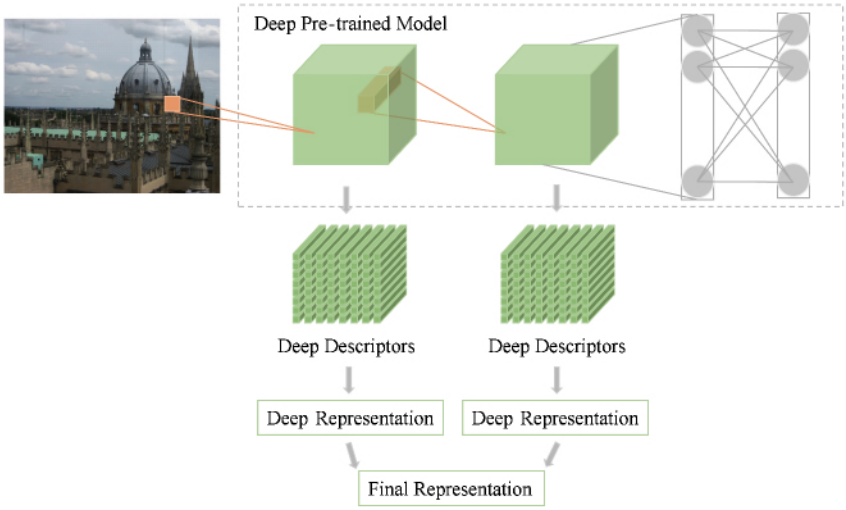

Fig. 5 Pipeline of image retrieval using multi-layer fusion

图 5 使用多层融合进行图像检索的流程图

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

然而,文献[17]发现简单将深度全连接特征级联起来性能不如只使用单一全连接特征.文献[51]认为这是因为我们不能给予所有层特征相同的重要性.

文献[51]采用了在线收集和选择[62]的方式自适应构造各层特征的权重,该文发现,单独使用低层特征性能不如单独使用高层特征,这是因为低层的纹理信息不具有足够强的判别能力.而多层特征融合取得了比单独使用特定层特征更好的性能.

SCDA 同时使用了 VGG-16 的 relu5-2 和 pool5 特征.SCDA 分别提取了 relu5-2 和 pool5 的图像表示向量,并将这两者加权级联作为最终表示向量.其中 pool5 表示的权重为 1,而 relu5-2 的权重为 0.5.SCDA 发现,relu5-2 特征的语义信息不如 pool5 特征丰富,但 relu5-2 特征对目标的检测比 pool5 更准.融合更低层的层,如 pool4,会对性能有些许的损失.此外,SCDA 同时使用了原图输入和原图水平翻转后的输入来提取深度特征,并级联为最终图像表示.

MS-RMAC[63]对 VGG-16 的 relu1-2,relu2-2,relu3-3,relu4-3 和 relu5-3 分别提取 R-MAC 特征,并将结果加权级联成最终图像表示.其中,各层权重是由免参 hedge 算法[64]得出,其基于各层的检索精度迭代调整各层权重.MS-RMAC 在实验中发现,relu1-2 和 relu2-2 对应权重为 0,relu3-3 对应权重不到 0.05,relu4-3 对应权重接近 0.2,而 relu5-3 对应权重接近 0.8.

6　实现细节

为了提高图像检索的性能,除了设计出更好的算法流程框架外,本节将简要介绍一些实现细节.通常情况下,将这些细节结合第 5 节介绍的算法流程,能提升图像检索性能,近年主流的无监督图像检索算法如 CroW①,Class weightied②,PWA③等均不同程度采用了 9 种实现细节:

1) 使用 ImageNet 数据集进行预训练.ImageNet 是一个经常被用于预训练模型的大规模数据集.ImageNet 利用 AMT 众包收集,其包含 120 万张训练图像和 5 万张验证图像以及 1 000 个类别.

2) VGGNet[65]是预训练模型的一个合适选择.VGGNet 是 ILSVRC 竞赛 2014 年目标竞赛任务的冠军和图像分类竞赛的亚军.由于 VGGNet 有良好的泛化性能,其在 ImageNet 上的预训练模型被广泛迁移到其他任务,包括图像检索.根据网络深度的不同,常用 VGGNet 模型有 VGG-16 和 VGG-19 两种,分别包含 16 和 19 个可学习参数的层(即卷积层或全连接层).

3) 使用原图大小输入.固定大小图像输入有时需要对图像做下采样,这会带来图像中的一些信息损失,而使用原图输入会保留更多的细节信息[51].我们也可以对图像做上采样,但要考虑后续特征提取过程中增加的计算开销.此外,保持原图长宽比也会带来更好的性能,这避免了使用固定长宽比带来的图像扭曲.

4) 使用经过 ReLU 激活函数之后的深度特征[18].pool5 特征是一个广泛使用的深度卷积特征,而 fc7 特征是一个广泛使用的深度全连接特征.

5) 使用求和汇合.当需要对深度特征进行聚合时,实践中发现,使用求和汇合通常比 BoW,VLAD,FV 等经典聚合方法更简单有效,也会取得比最大汇合及更好的性能[19,21].

6) 对提取得到的图像表示进行 ℓ_2 规范化,PCA 白化和再次 ℓ_2 规范化[66].PCA 白化是用于应对图像表示各维之间的共现(co-occurrence)效应[67],并对高维的表示向量进行降维.最后一次 ℓ_2 规范化是为了使表示向量的空间由欧氏空间映射到余弦空间.具体来说,给定 2 个向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$,其欧氏距离和余弦相似性分别定义为

$$\|\boldsymbol{u}-\boldsymbol{v}\|=\sqrt{-2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}+\|\boldsymbol{u}\|^2+\|\boldsymbol{v}\|^2},$$

$$\cos\langle\boldsymbol{u},\boldsymbol{v}\rangle=\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|},$$

当 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 经过 ℓ_2 规范化后,即 $\bar{\boldsymbol{u}}=\boldsymbol{u}/\|\boldsymbol{u}\|$,$\bar{\boldsymbol{v}}=\boldsymbol{v}/\|\boldsymbol{v}\|$,此时:

$$\|\bar{\boldsymbol{u}}-\bar{\boldsymbol{v}}\|=\sqrt{-2\bar{\boldsymbol{u}}^{\mathrm{T}}\bar{\boldsymbol{v}}},$$

$$\cos\langle\bar{\boldsymbol{u}},\bar{\boldsymbol{v}}\rangle=\bar{\boldsymbol{u}}^{\mathrm{T}}\bar{\boldsymbol{v}},$$

它们将给出相同的相似度排序.也就是说,经过 ℓ_2 规范化后的欧氏距离检索结果实质是在原空间利用余弦相似性的检索结果.

7) 从另一个独立的数据集学习 PCA 的参数.由于深度图像表示相比经典图像表示的维数通常更

① CroW:https://github.com/yahoo/crow/

② Class-weighted convolutional features:https://github.com/imatge-upc/retrieval-2017-cam/

③ PWA:https://github.com/XJhaoren/PWA/

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

低,学习 PCA 参数的过拟合风险更高[18-19]. 因此,实际应用中通常使用另一个相似数据集学习其 PCA 参数. 通常做法是,使用 Oxford 数据集学习 Paris 的 PCA 参数,反之亦然,并使用 Flickr100k 学习 Holidays 的参数.

8) 幂律规范化[30,57]有时会提升性能. 幂律规范化的提出是为了解决图像中的"视觉爆发"现象. 常见处理方法是对图像表示做带符号开根号:

$$y=\mathrm{sign}(x)\sqrt{|x|},$$

之后再对 y 做 ℓ_2 规范化.

9) 对检索得到的排序进行后处理. 在查询图像表示对数据库图像表示进行最近邻搜索之后,将得到一个对数据库图像从最相似到最不相似的排序. 我们可以进行后处理,对排序结果进行微调. 例如查询展开(query expansion)[68]是将排序最前的几个(如 10 个)结果的表示向量求和汇合并 ℓ_2 规范化,将其作为新的查询向量重新进行检索. 这通常会得到 2%~3%的 mAP 提升. 此外,空间重排(spatial reranking)[28]也是常用的后处理技术,其旨在对图像局部进行更细的匹配.

7 各算法性能比较

不同图像检索算法在 Oxford5k, Paris6k, Oxford105k, Paris106k 和 Holidays 数据集上的性能比较见表 1. 为公平比较,这些结果均没有使用后处理技术. 由于 Ukbench 数据集近年来使用相对较少,所以在表 1 中没有列出.

Table 1 Comparisons of Different Image Retrieval Approaches

表 1 不同图像检索算法性能比较

Method	Model	Dimension	Oxford5k/%	Paris6k/%	Oxford105k/%	Paris106k/%	Holidays/%
Triangulation Embedding[33]	SIFT	8 192	67.6		61.1		77.1
FAemb[36]	SIFT	16 384	70.9				78.7
RVD-W[37]	SIFT	16 384	68.9		66.0		78.8
Neural code[17]	AlexNet	512	43.5		39.2		
CNN-SL[38]	AlexNet	4 096	41.7	58.1			
MOP-CNN[18]	AlexNet	2048					80.2
Off-the-shelf[39]	Overfeat	Not fixed	68.0	79.5			84.3
Patch-CKN[42]	Patch-CKN	~65 000	56.5				79.3
OLDFP[45]	AlexNet	512	59.0	63.0			80.1
CCS[48]	GoogLeNet	512	67.3	72.2			
SPoC[19]	VGG-19	256	53.1		50.1		80.2
R-MAC[20]	VGG-16	512	66.9	83.0	61.6	75.7	
CroW[21]	VGG-16	512	70.8	79.7	65.3	72.2	85.1
uCroW[21]	VGG-16	512	69.7	78.6	64.1	71.0	83.9
OnA of Ref [22]	VGG-16	512	73.6	85.5			
OfA of Ref [22]	VGG-16	512	71.2	80.5	67.2	73.3	
PWA[23]	VGG-16	4 096	**79.1**	**86.1**	**73.6**	**80.4**	
Visual instance[50]	VGG-19	512	46.2	67.4			74.6
VLAD-CNN[52]	GoogLeNet	128	55.8	58.3			83.6
BLCF[53]	VGG-16	~25 000	73.8	82.0	59.3	64.8	
SCDA[54]	VGG-16	512	67.7				**92.1**
SWFV[58]	VGG-16	512	68.5	82.2			
TSWFV[58]	VGG-16	512	68.8	83.0			
Good practice[51]	VGG-16	9 664	71.3				84.2
MS-RMAC[63]	VGG-16	1 472	68.9	77.6			86.7

Note: The state-of-the-art results are indicated in boldface.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表1由上至下分为5个部分:基于手工特征、基于深度全局特征、基于局部表示聚合、基于深度卷积特征聚合和基于多层融合的图像检索算法.可以看出,目前的研究热点是基于深度局部特征的图像检索,尤其在于基于深度卷积特征聚合的图像检索算法.这类算法也取得了目前最好的效果.

8　未来研究方向

深度图像检索的研究方兴未艾,亟待后续研究的进行.本节对图像检索的未来研究方向进行展望.

1)更有效地利用深度卷积特征.影响图像检索性能的关键是提取得到的图像表示的质量,而深度卷积特征十分高维且稀疏,具有判别能力的信息隐藏在深度卷积特征中.如何更有效地利用这些深度卷积特征,将成为未来研究的一大突破点.

2)多层融合.局部特征相比全局特征对图像的平移、遮挡等更不敏感,通常会取得更好的性能.而另一方面,由于局部特征不包含图像的全局信息,因此,基于局部特征的检索将可能返回局部相似的无关图像.

使用全局特征抑或是局部特征不应是互斥的关系.由于深度卷积神经网络不同层特征具有层次性,不同层特征的语义信息可以相互补充.此外,多层融合也可以看作是一种集成学习.并且和经典的多模型集成方法相比,这种集成只需要网络前馈一次,十分高效.

然而,目前在深度特征多层融合方面的研究还较少,其中一个原因是对多层融合应该选择哪些层、如何做融合这些方面缺乏合适的指导.而且,并不是融合的层越多效果越好,有时甚至会起到相反作用.

3)特定应用场景下的图像检索.本文所介绍的算法均属于通用图像检索算法,即算法不对图像中目标的内容和性质做出假设.另一方面,我们可以利用图像中目标的性质以设计针对特定应用场景下的图像检索算法,例如多标记图像检索[69]、基于草图的图像检索[70]、医学CT图像检索[71]、细粒度图像检索[54,72]、场景检索[73]、行人检索[74]、车辆检索[75]、图标检索[76]、商品检索[77]、人脸检索[78]等.

通用图像检索算法的应用范围更广,但通用图像检索研究和特定应用下的图像检索研究两者不应该是独立的.通用图像检索算法中的设计思想也可以被应用于特定应用下图像检索.

4)检索效率.由于图像检索要面临非常庞大的数据库图像,因此图像检索算法效率对该算法能否实际应用至关重要.这包括提取图像表示的效率和对查询图像进行检索的效率.

目前,基于深度卷积特征聚合的方法比基于局部表示聚合的方法有更广的应用,其中一个原因即是基于深度卷积特征聚合的方法更加高效.另一方面,乘积量化(product quantization)[79]和Hash[34]是常用的快速检索算法.

虽然目前的研究工作对检索性能更加看重,但检索效率仍是不容忽视的重要方面.基于深度特征的方法近年来取得了突出进展,然而,深度学习模型需要占用大量的与计算相关的资源.因此目前深度学习领域的一个热点是研究资源受限的深度学习[80].

5)更大更高质量更通用的标准数据集.目前图像检索研究中主流使用的标准数据集,尽管可供选择的余地不小,但都存在一个共同的不足之处:规模还比较小,检索内容也比较单一(如集中在建筑物、风景、室内物体中).而图像检索是一个与实际应用密切相关的研究领域,若想使图像检索在实际场景中得到广泛应用,就不得不考虑诸如光照、模糊、遮挡、低分辨率、物体干扰等复杂场景下的图像检索问题.因此,构建更大、更高质量、更通用的标准数据集成为了未来亟需解决的一大问题.

9　总　　结

图像检索是计算机视觉领域一个重要的研究方向,而深度特征的出现为其带来了新的发展机遇.本文对基于经典手工特征的图像检索方法做简要回顾,并从深度全局特征和深度局部特征2个角度,对近年来基于深度特征的图像检索的代表算法予以综述.并介绍了常用数据集和对该领域未来可能的发展机遇进行展望.

参　考　文　献

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

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

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[3] Liu Ying, Zhang Dengsheng, Lu Guojun, et al. A survey of content-based image retrieval with high-level semantics [J]. Pattern Recognition, 2007, 40(1): 262-282

[4] Zheng Liang, Yang Yi, Tian Qi. SIFT meets CNN: A decade survey of instance retrieval [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1224-1244

[5] Mukherjee A. Intelligent Analysis of Multimedia Information [M]. Hershey: IGI Global, 2017: 143-180

[6] Jain A K, Vailaya A. Image retrieval using color and shape [J]. Pattern Recognition, 1996, 29(8): 1233-1244

[7] Manjunath B S, Ma Weiying. Texture features for browsing and retrieval of image data [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1996, 18(8): 837-842

[8] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope [J]. International Journal of Computer Vision, 2001, 42(3): 145-175

[9] Oliva A, Torralba A. Scene-centered description from spatial envelope properties [G] //LNCS 2525: Proc of the Biologically Motivated Computer Vision Second Int Workshop. Berlin: Springer, 2002: 263-272

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

[11] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos [C] //Proc of the 9th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2003: 1470-1477

[12] Qiu Guoping. Indexing chromatic and achromatic patterns or content-based colour image retrieval [J]. Pattern Recognition, 2002, 35(8): 1675-1686

[13] Nister D, Stewenius H. Scalable recognition with a vocabulary tree [C] //Proc of the 2006 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2006: 2161-2168

[14] Perronnin F, Dance C R. Fisher kernels on visual vocabularies for image categorization [C] //Proc of the 2007 IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2007

[15] Perronnin F, Sánchez J, Mensink T. Improving the Fisher kernel for large-scale image classification [G] //LNCS 9905: Proc of the 11th European Conf on Computer Vision. Berlin: Springer, 2010: 143-156

[16] Jégou H, Perronnin F, Douze M, et al. Aggregating local descriptors into compact codes [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 34(9): 1704-1716

[17] Babenko A, Slesarev A, Chigorin A, et al. Neural codes for image retrieval [G] //LNCS 8689: Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 584-599

[18] Gong Yunchao, Wang Liwei, Guo Ruiqi, et al. Multi-scale orderless pooling of deep convolutional activation features [G] //LNCS 8689: Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 392-407

[19] Babenko A, Lempitsky V S. Aggregating local deep features for image retrieval [C] //Proc of the 2015 IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 1269-1277

[20] Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations [C] //Proc of the Int Conf on Learning Representations. 2016

[21] Kalantidis Y, Mellina C, Osindero S. Cross-dimensional weighting for aggregated deep convolutional features [G] //LNCS 9905: Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 685-701

[22] Jiménez A, Alvarez J M, Giro-i-Nieto X. Class-weighted convolutional features for visual instance search [C] //Proc of the British Machine Vision Conference. Durham: British Machine Vision Association, 2017

[23] Xu Jian, Shi Cunzhao, Qi Chengzuo, et al. Unsupervised part-based weighting aggregation of deep convolutional features for image retrieval [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto: AAAI, 2018: 7436-7443

[24] Russakovsky O, Deng Jia, Su Hao, et al. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252

[25] Donahue J, Jia Yangqing, Vinyals O, et al. DeCAF: A deep convolutional activation feature for generic visual recognition [C] //Proc of the 31st Int Conf on Machine Learning. New York: ACM, 2014: 647-655

[26] Cimpoi M, Maji S, Kokkinos I, et al. Deep filter banks for texture recognition and segmentation [J]. International Journal of Computer Vision, 2016, 118(1): 65-94

[27] Ghodrati A, Diba A, Pedersoli M, et al. DeepProposals: Hunting objects by cascading deep convolutional layers [J]. International Journal of Computer Vision, 2017, 124(2): 115-131

[28] Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching [C] //Proc of the 2007 IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2007

[29] Philbin J, Chum O, Isard M, et al. Lost in quantization: Improving particular object retrieval in large scale image databases [C] //Proc of the 2008 IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2008

[30] Jégou 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

[31] Bay H, Tuytelaars T, Gool L J V. SURF: Speeded up robust features [G] //LNCS 3951: Proc of the 9th European Conf on Computer Vision. Berlin: Springer, 2006: 404-417

[32] Harris Z S. Distributional structure [J]. Word, 1954, 10(2/3): 146-162

[33] Jégou H, Zisserman A. Triangulation embedding and democratic aggregation for image search [C] //Proc of the 2014 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 3310-3317

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[34] Perronnin F, Liu Yan, Sanchez J, et al. Large-scale image retrieval with compressed Fisher vectors [C] //Proc of the 23rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 3384-3391

[35] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] // Proc of Advances in Neural Information Processing Systems 25. Cambridge, MA: MIT Press, 2012: 1106-1114

[36] Do T T, Tran Q D, Cheung N M. FAemb: A function approximation-based embedding method for image retrieval [C] //Proc of the 2015 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3556-3564

[37] Husain S S, Bober M. Improving large-scale image retrieval through robust aggregation of local descriptors [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 39(9): 1783-1796

[38] Wan Ji, Wang Dayong, Hoi S C H, et al. Deep learning for content-based image retrieval: A comprehensive study [C] // Proc of the ACM Int Conf on Multimedia. New York: ACM, 2014: 157-166

[39] Razavian A S, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: An astounding baseline for recognition [C] // Proc of the 2014 IEEE Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2014: 512-519

[40] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories [C] //Proc of the 2006 IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2006: 2169-2178

[41] Mikolajczyk K, Tuytelaars T, Schmid C, et al. A comparison of affine region detectors [J]. International Journal of Computer Vision, 2005, 65(1-2): 43-72

[42] Paulin M, Douze M, Harchaoui Z, et al. Local convolutional features with unsupervised training for image retrieval [C] // Proc of the 2015 IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 91-99

[43] Mikolajczyk K, Schmid C. Scale & affine invariant interest point detectors [J]. International Journal of Computer Vision, 2004, 60(1): 63-86

[44] Mairal J, Koniusz P, Harchaoui Z, et al. Convolutional kernel networks [C] //Proc of Advances in Neural Information Processing Systems 27. Cambridge, MA: MIT Press, 2014: 2627-2635

[45] Mopur K R, Babu R V. Object level deep feature pooling for compact image representation [C] //Proc of the 2015 IEEE Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2015: 62-70

[46] Uijlings J, van de Sande K, Gevers T, et al. Selective search for object recognition [J]. International Journal of Computer Vision, 2013, 104(2): 154-171

[47] Boureau Y L, Bach F R, Lecun Y, et al. Learning mid-level features for recognition [C] //Proc of the 23rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 2559-2566

[48] Yan Ke, Wang Yaowei, Liang Dawei, et al. CNN vs. SIFT for image retrieval: Alternative or complementary? [C] // Proc of the 2016 ACM Conf on Multimedia Conf. New York: ACM, 2016: 407-411

[49] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges [G] //LNCS 8689: Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 391-405

[50] Razavian A S, Sullivan J, Maki A, et al. Visual instance retrieval with deep convolutional networks [J]. ITE Trans on Media Technology and Applications, 2016, 4(3): 251-258

[51] Zheng Liang, Zhao Yali, Wang Shengjin, et al. Good practice in CNN feature transfer [OL]. [2017-12-30]. https://arxiv.org/abs/1604.00133v1

[52] Ng J Y-H, Yang Fan, Davis L S. Exploiting local features from deep networks for image retrieval [C] //Proc of the 2015 IEEE Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2015: 53-61

[53] Mohedano E, McGuinness K, O'Connor N E, et al. Bags of local convolutional features for scalable instance search [C] // Proc of the 2016 ACM on Int Conf on Multimedia Retrieval. New York: ACM, 2016: 327-331

[54] Wei Xiu-Shen, Luo Jian-Hao, Wu Jianxin, et al. Selective convolutional descriptor aggregation for fine-grained image retrieval [J]. IEEE Trans on Image Processing, 2017, 26(6): 2868-2881

[55] Hinton G E. Learning distributed representations of concepts [C] //Proc of the Annual Conf of the Cognitive Science Society. Boston: Cognitive Science Society, 1986

[56] Bengio Y, Courville A C, Vincent P. Representation learning: A review and new perspectives [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828

[57] Jégou H, Douze M, Schmid C. On the burstiness of visual elements [C] //Proc of the 2009 IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 1169-1176

[58] Qi Chengzuo, Shi Cunzhao, Xu Jian, et al. Spatial weighted Fisher vector for image retrieval [C] //Proc of the 2017 IEEE Int Conf on Multimedia and Expo. Piscataway, NJ: IEEE, 2017: 463-468

[59] Zhou Bolei, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization [C] //Proc of the 2016 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2921-2929

[60] Hariharan B, Arbeláez P A, Girshick R B, et al. Hypercolumns for object segmentation and fine-grained localization [C] //Proc of the 2015 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 447-456

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[61] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651

[62] Zheng Liang, Wang Shengjin, Tian Lu, et al. Query-adaptive late fusion for image search and person re-identification [C] //Proc of the 2015 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1741-1750

[63] Li Yang, Xu Yulong, Wang Jiabao, et al. MS-RMAC: Multiscale regional maximum activation of convolutions for image retrieval [J]. IEEE Signal Processing Letters, 2017, 24(5): 609-613

[64] Chaudhuri K, Freund Y, Hsu D J. A parameter-free hedging algorithm [C] //Proc of Advances in Neural Information Processing Systems 22, Cambridge, MA: MIT, 2009: 297-305

[65] Simonyan K, Zisserman A: Very deep convolutional networks for large-scale image recognition [C] //Proc of the Int Conf on Learning Representations. 2014 [2018-01-24]. https://arxiv.org/abs/1409.1556

[66] Jégou H, Chum O. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening [G] // LNCS 7572: Proc of the 12th European Conf on Computer Vision. Berlin: Springer, 2012: 774-787

[67] Chum O, Matas J. Unsupervised discovery of co-occurrence in sparse high dimensional data [C] //Proc of the 23rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 3416-3423

[68] Chum O, Philbin J, Sivic J, et al. Total recall: Automatic query expansion with a generative feature model for object retrieval [C] //Proc of the 11th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2007: 1-8

[69] Lai Hanjiang, Yan Pan, Shu Xiangbo, et al, Instance-aware hashing for multi-label image retrieval [J]. IEEE Trans on Image Processomg, 2016, 25(6): 2469-2479

[70] Qian Xueming, Tan Xianglong, Zhang Yuting, et al. Enhancing sketch-based image retrieval by re-ranking and relevance feedback [J]. IEEE Trans on Image Processing, 2016, 25(1): 195-208

[71] Dubey S R, Singh S K, Singh R K. Local wavelet pattern: A new feature descriptor for image retrieval in medical CT databases [J]. IEEE Trans on Image Processing, 2015, 24(12): 5892-5903

[72] Xie Lingxi, Wang Jingdong, Zhang Bo, et al. Fine-grained image search [J]. IEEE Trans on Multimedia, 2015, 17(5): 636-647

[73] Torii A, Arandjelovic R, Sivic J, et al. 24/7 place recognition by view synthesis [C] //Proc of the 2015 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1808-1817

[74] Gong Shaogang, Tao Xiang. Visual Analysis of Behaviour [M]. Berlin: Springer, 2014: 301-313

[75] Liu Xinchen, Liu Wu, Ma Huadong, et al. Large-scale vehicle re-identification in urban surveillance videos [C] // Proc of the 2016 IEEE Int Conf on Multimedia and Expo. Piscataway, NJ: IEEE, 2016: 1-6

[76] Revaud J, Douze M, Schmid C. Correlation-based burstiness for logo retrieval [C] //Proc of the 20th ACM Int Conf on Multimedia. New York: ACM, 2012: 965-968

[77] Zhou Ye, Zhang Junping. Multi-scale deep learning for product image search [J]. Journal of Computer Research and Development, 2017, 54(8): 1824-1832 (in Chinese)
(**周晔, 张军平. 基于多尺度深度学习的商品图像检索**[J]. **计算机研究与发展**, 2017, 54(8): 1824-1832)

[78] Desai R, Sonawane B. Gist, HOG, and DWT-based content-based image retrieval for facial images [C] //Proc of the Int Conf on Data Engineering and Communication Technology. Berlin: Springer, 2017: 297-307

[79] Jegou H, Douze M, Schmid C. Product quantization for nearest neighbor search [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2011, 33(1): 117-128

[80] Wu Jianxin, Gao Binbin, Wei Xiushen, et al. Resource-constrained deep learning: Chanllenges and practices [J]. Scientia Sinica Informationis, 2018, 48: 501-510 (in Chinese)
(**吴建鑫, 高斌斌, 魏秀参, 等. 资源受限的深度学习: 挑战与实践**[J]. **中国科学: 信息科学**, 2018, 48: 501-510)

Zhang Hao, born in 1994. Master candidate. His main research interests include computer vision and machine learning.

Wu Jianxin, born in 1978. PhD, professor, PhD supervisor. His main research interests include computer vision and machine learning.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net