多媒体工程：2016

——图像检索研究进展与发展趋势

于俊清[1]，吴泽斌[1]，吴飞[2]，孙立峰[3]

1. 华中科技大学计算机科学与技术学院，武汉 430074；2. 浙江大学计算机学院，杭州 310058；
3.清华大学计算机科学与技术系，北京 100084

摘要：目的 基于内容的图像检索方法利用从图像提取的特征进行检索，以较小的时空开销尽可能准确的找到与查询图片相似的图片。**方法** 本报告从浅层特征、深层特征和特征融合三个方面对图像检索国内外研究进展和面临的挑战进行介绍，并对未来的发展趋势进行展望。**结果** SIFT 存在缺乏空间几何信息和颜色信息，高层语义的表达不够等问题；而 CNN（Convolutional Neural Network,卷积神经网络）特征则往往缺乏足够的底层信息。为了丰富描述符的信息，通常将 SIFT 与 CNN 等特征进行融合。融合方式主要包括：串连、核融合、图融合、索引层次融合和得分层(Score-level）融合。“融合”可以有效的利用不同特征的互补性，提高检索的准确率。**结论** 与 SIFT 相比，CNN 特征的通用性及几何不变性都不够强，依然是图像检索领域面临的挑战。

关键词：尺度不变特征，卷积神经网络，特征融合，图像检索

Research on multimedia technology 2016——advances and trend of image retrieval

Yu Junqing[1], Wu Zebin[1], Wu Fei[2], Sun Lifeng[3]

1. Computer Department of HuaZhong University of Science and Technology, Wuhan 430074,China;

2. Computer Department of Zhejiang University, Hangzhou 310058,China;

3. Computer Department of Tsinghua University, Beijing 100084,China.

Abstract: Objective Content-based image retrieval uses feature extracted from the image to retrieve similar images accurately from a large-scale dataset with small memory and time consumption. SIFT (Scale Invariant Feature Transform) feature is robust to translation, scaling, rotation, viewpoint changing and occlusion, and is fast to be extracted, making it widely used theoretically and practically. But some shortcomings exist in SIFT: lacking spatial geometric information; lacking color information. Convolutional Neural Network (CNN) has good domain transferability and the deep feature from pre-trained CNN can be applied to various domains.CNN deep feature

收稿日期：2017-；**修回日期：**2017-
基金项目：国家自然科学基金（61572211）
第一作者简介：于俊清（1975-），男，教授，博士生导师，2002 年于武汉大学计算机学院获工学博士学位，主要研究领域为多媒体信息处理与检索、多核计算与数据编译、网络安全与教育信息化等。E-mail: yjqing@hust.edu.cn
Supported by: National Natural Science Foundation of China（61572211）

© 中国图象图形学报版权所有

recently has attracted much attention and shown superior performance over SIFT. However, contrary to the shortcoming of SIFT, CNN features lack shallow information. So, SIFT is usually fused with CNN features as well as other shallow features. **Method** This report reviews the recent advances and challenges with image retrieval in the world and in China, including shallow feature, deep feature and feature fusion. The future development trend is also prospected. For the shallow features, we mainly reviews SIFT and its variants, the encoding methods and the development of the these methods. For deep feature, we divide the descriptors of the features into different categories according to the type of CNN layer used: fully-connected layer, convolutional layer, softmax-layer. Many features can be extracted from a convolutional layer and many pooling methods are proposed. **Result** The encoding methods of SIFT mainly including BOF (Bag Of Feature), VLAD (Vector of Locally Aggregated Vectors), FV (Fisher Vectors), TE (Triangulation Embedding), and mostly consist of two steps: embedding and aggregation (or pooling). For CNN features, features from the fully connected layer of CNN is usually used for its good transferability and accuracy, however, the deep features from the convolutional layer recently become more and more attractive, because the convolutional features can be effectively combined with a variety of pooling methods like sum-pooling, max-pooling, VLAD-pooling, FV-pooling, and show good performance in the domains of image classification and retrieval. The fusion methods can mainly be divided into five types: concatenation, kernel fusion, graph fusion, index-level fusion, and score-level fusion. Concatenation, kernel fusion and index-level fusion work directly on different features, but graph fusion and score-level fusion work on the retrieval results of different features. Fusion makes use of the complementary of different features and can improves the accuracy of image retrieval effectively. **Conclusion** SIFT and CNN feature is complementary to each other: SIFT contains rich low-level information, and CNN features contain rich high semantic information; SIFT has a good property of invariance which is the shortcoming of CNN features. Fusion is a effective way to make most use of the image information, however, the time and space consumption will inevitably increase and a good algorithm to distinguish the good feature from the bad is still to be studied. For the present, the generalizability and geometric invariance of CNN feature are inferior to SIFT and this is still a challenge for image retrieval researchers. The generalizalibily of CNN features is limited by the domain and statistic difference between the source task (usually ImageNet) and the target task. Finetuing is a good strategy to solve this problem, however, finetuning needs extra labeled dataset similar to the target task. To enhance the geometric invariance of CNN, the CNN descriptor space-consumption and extracting time will inevitably increase and usually only the scale invariance is considered for simplicity, ignoring the other aspects of invariance. Besides, the number of CNN feature from one image is usually much smaller than SIFT and can’t capture enough information for encoding. The most usually used CNNs are designed for image classification tasks, not for image retrieval, however, image retrieval is a more fine-grained domain: the algorithm needs to find the similar images, not just the images from one class. So a CNN trained for image retrieval may be a good future direction. More work is still need to be done to strike a better balance among generalizability, invariance, memory-consumption and extracting time for a effective and efficient image retrieval descriptor.

Keywords: Scale Invariant Feature Transform, Convolutional Neural Network, Feature fusion, Image retrieval

0 引 言

随着计算机网络、社交媒体、数字电视和多媒体获取设备的快速发展，以图像和视频为代表的多媒体数据的生成、处理和获取变得越来越方便，多媒体应用日益广泛，数据量呈现出爆炸性的增长，已经成为大数据时代的主要数据对象。如何在海量的图像大数据中以较小的时空开销准确地找到一幅感兴趣的图像，已经成为近年来多媒体和信息检索领域的重要研究热点。

© 中国图象图形学报版权所有

基于内容的图像检索(CBIR)方法利用从图像提取的特征来进行检索。常用的图像特征主要有颜色、纹理和形状，包括局部特征和全局特征。局部特征是基于图像的某个区域提取的图像描述符，如尺度不变特征(SIFT)[1-2]。全局描述符基于整幅图像提取的描述符，如 GIST[3] 。全局特征对图像的压缩率较高，但区分力不强；局部特征的区分力强，但数目太多，故而各种编码方法被提了出来，如特征袋(BOF)[4] ，Fisher 向量 (FV)[5] ，以及 VLAD (Vector of Locally Aggregated Descriptors)[6] 等。BOF，VLAD，FV 等描述符通常继承了局部特征的部分不变性，如对平移、旋转、缩放、光照和遮挡等与语义相关不大的因素保持不变。

基于 SIFT 等图像描述符的检索效果相对于现有的其他特征明显改进，然而，SIFT 存在如下几个问题：1)缺乏空间几何信息; 2)缺乏颜色信息;3）缺乏高层语义。为了丰富描述符的信息，通常将 SIFT 与其他的特征进行融合。如文献[7] 中，利用核来融合多种特征，形成语义属性特征，再与 FV 相串联以融合 SIFT 特征。文献[8] 则是通过图来融合 SIFT 与颜色特征，以提高检索的准确率。文献[9] 则是通过一个 2 维索引结构来融合 SIFT 与颜色特征。

SIFT 描述的是图像的底层特征，无法很好地表示图像的高层语义，因此，基于数据驱动的图像特征提取方法被提出，神经网络就是其中之一。然而，最初之时，神经网络的层与层之间是全连接的，参数太多，网络不能太深，否则，训练将非常困难。一方面，训练将非常耗时；另一方面，当时没有带标签的大数据集，训练网络时容易发生过拟合。卷积神经网络(CNN) [10-12] 的神经元是局部连接的，是一种易于训练的网络，这使得 CNN 可以更深。随着 ImageNet [13] 等带标签的大数据集的提出，CNN 得以广泛应用。在 ILSVRC2012 (imagenet large scale visual recognition challenge 2012) 比赛上，Alex 等人[14] 提出的 CNN 框架取了冠军，远远超过了前人的结果，此 CNN 通常被称为 AlexNet。AlexNet 拥有 5 个卷积层，3 个全连接层，6 000 万参数，65 万个神经元。为了加快训练，AlexNet 以 ReLU[15] 作为激励单元，利用 GPU 进行加速；此外，AlexNet 在全连接层使用了 Dropout[16] -17]，以减弱过拟合现象。继 AlexNet 之后，VGGNet（或 OxfordNet)[18] 、GoogLeNet[19] (ILSVRC2014 冠军)、ResNet[20] (ILSVRC2015 冠军) 等新的 CNN 框架相继被提出，CNN 被广泛应用到图像分类[14,21]、语义分割[22] 、动作识别[23-24]、语音识别[25-26]和机器翻译[27-28]等领域，并几乎都获得了当时最好的结果。

文献[29-31]等表明，CNN 具有良好的跨域特性（或通用性），从预训练的 CNN 提取的特征可以被广泛应用到各个领域的各种数据集。文献[31] 中，预训练的 CNN 被用于图像分类、属性检测、细粒度识别、图像检索，均取得了优良的结果。文献[31] 表明，源任务数据集(ImageNet)与目标任务数据集的差异越小，视觉识别任务的效果越好。基于卷积神经网络的深度学习得到的特征不仅保持了一定的不变性，而且还包含了更多的高层语义信息，可以有效地缩小底层特征与高层语义之间的鸿沟[32] -[33] 。CNN 全连接层特征性能较好，是最常使用的 CNN 特征，然而，CNN 全连接层特征的几何不变性无法与 SIFT 相比，且缺乏对局部细节的描述，因而卷积层特征也成为了研究的热点。

本报告将从浅层特征、深层特征和特征融合 3 个方面对国内外研究进展和面临的挑战进行介绍，并对未来的发展趋势进行展望。

1 国际研究现状

1.1 浅层特征

浅层特征提取方法基于领域知识通过固定的算法提取特征，目前被广泛采用且效果良好的浅层特征有 SIFT, GIST 等。根据在提取过程中使用的是图像区域还是整幅图像，这些特征又可分为局部特征（如 SIFT）和全局特征（如 GIST）。全局特征对图像信息的压缩率更高，但检索的准确率不如局部特征。本节首先描述了 SIFT 特征及其各种变体，继而介绍了对 SIFT 编码形成的各种浅层描述符。

SIFT[1] 由 David 于 1999 年提出，用于解决图

© 中国图象图形学报版权所有

像匹配这个计算机视觉领域的基本问题。SIFT 不仅对尺度、旋转和平移具有不变性，而且对遮挡、噪声及光照变化也具有较好的鲁棒性；SIFT 的生成过程非常快，可以满足实时性的要求。SIFT 被广泛用于对象和场景识别，而其潜在应用则不胜枚举，如 3D 重建、动作跟踪、机器人定位及图像全景缝合等。Ke 等人在 SIFT 的基础上提出了 PCA-SIFT[34]。PCA-SIFT 在 SIFT 提取算法中使用主成分分析(PCA)，得到的是 36 维的特征向量，在图像检索方面取得了比 SIFT 更高的准确率与速度。Mikolajczyk 等人[35] 对 SIFT，PCA-SIFT 等 10 种描述符在使用各种区域检测子的情况下进行比较，提出了 SIFT 描述符的一种变体，梯度—位置—方向直方图(GLOH)，以提高描述符的区分性和鲁棒性。不同于基于差分图像的 SIFT，Bay 等人[36] 基于积分图像对 SIFT 进行全面改进，提出了一个 SIFT 的加速版，加速的健壮性特征 (SURF)，SURF 的稳定点检测器以及区域描述器都与 SIFT 的不同。Arandjelovic 等人[37] 利用 1-范数归一化和平方根变换得到了比 SIFT 区分力更强的描述符 RootSIFT。

由于一幅图像中包含的局部特征数目不同且数量众多，可达数百甚至数千，而每个局部特征的维度较高，因此不方便图像间的快速相似度比较，无法适应大规模图像检索对存储开销和快速响应的要求。针对上述问题，研究者们提出了多种特征编码方法，可以将不同图像中数目不一的一组局部特征变换生成一个固定长度的特征表达，以实现高效图像检索。常见的特征编码方法包括 BOF, FV, VLAD 等。

1.1.1 BOF

BOF[4] 在 2003 年由 Google Robotics Research Group 的研究人员提出，这一概念是文本检索在图像检索领域的推广。在文本检索中，一篇文章被看成是词的集合，每篇文章对应一个词频向量，将此概念推广到图像检索中，一幅图像被看成一篇文章，由若干个“视觉单词”构成，每幅图像对应一个 tf-idf（词频-逆文档频率）[38] 向量。“视觉单词”就是由 k-means 算法对图像的视点不变区域提取的 SIFT 向量聚类生成的中心向量，也称为码字，所有“视觉单词”的集合称为“词典”或“码书”。BOF 的码书通常较大，维度较高，是一个比较稀疏的向量，可以利用向量量化方法来形成倒排索引，以提高检索速度，检索结果按查询图像与数据库图像的 tf-idf 向量间的余弦距离排序。为了减少 BOF 对空间的占用，以适合大数据集，法国信息与自动化研究所 INRIA 的 Jegou 等人[39] 提出了对 BOF 进行二值化并对其进行稀疏压缩的方案，在 BOF 的基础上提出了一种二值化 BOF－miniBOF[39]。miniBOF 占用的空间比 BOF 减少了至少一个数量级，检索准确率没有明显的下降。

1.1.2 VLAD

法国 Jegou 等人[6] 针对图像检索应用，联合优化内存占用、检索准确率与检索速度，提出了 VLAD。VLAD 对 k-means 生成的每一个 Voronoi-cell 中的特征的残差求和，形成一个子向量，然后将这些子向量串连起来。为了减少 VLAD 对空间的占用，可以对 VLAD 进行 PCA 变换，为了加快使用倒排列表对图像的检索，可以对 VLAD 用积量化 (PQ)[40] 编码和非对称距离计算（ADC)。为了保证各个分量方差的均衡，可在 PCA 后再进行一个正交变换。在 INRIA Holidays、UKB 等数据集上的实验结果表明，VLAD 的测试准度率(或区分力）要高于 BOF、FV 和 miniBOF。Arandjelovic 等人[41] 进一步提出了内部归一化（intra-normalization），以消除“视觉 爆发(visual burstiness)”现象[42]；为了应对数据库的变动问题，Arandjelovic 等人还提出了一个码书更新策略。

Wang 等人结合 SIFT 特征的角度信息提出了 gVLAD(geometric VLAD)[43]，比普通的 VLAD 的检索准确率有了较大的提高，取得了优于 Intra-VLAD[41] 的性能。Wang 等人研究了小数据集高维和大数据集(Holidays1M)低维(D=128)两种情况下的 gVLAD 性能，表明了 gVLAD 的优越性。

© 中国图象图形学报版权所有

1.1.3 FV

BOF 要得到较高的准确率，通常需要较大的码书（几万至几十万），时空开销较大，训练也较为困难。Perronnin 等人[5] [44] 利用一个混合高斯模型来近似底层特征向量(如 SIFT)的分布，生成分类用的码书，并用这个模型参数（权值、均值和标准差）的梯度来作为图像的一个新的表示方法，并将这个梯度向量用于分类。这个“梯度向量”就是 FV 向量。BOF 和 VLAD 本质上是 FV 的特征情形：BOF 仅使用了频数信息（0 阶统计信息），VLAD 仅使用了均值梯度信息（1 阶统计信息）。VLAD 可以看作是 FV 的硬分配版本：用 k-means 代替了 GMM。Perronnin 等人表明，仅使用均值与标准差的梯度信息即已达到最优，故而通常不用权值梯度信息。FV 以较小的码书（几十至几百）取得了优于 BOW 的性能。Perronnin 等人对 FV 进一步改进，提出了 IFV(improved FV)[45] 。IFV 利用了 2-范数归一化、幂律标准化（power-normalization）和空间金字塔（spatial pyramid）[46] 3 种策略。“幂律标准化”本质上用于消除“视觉爆发”现象，而空间金字塔则是提供了多个尺度的信息。Perronnin 等人[47] 进一步研究了 FV 的二值化压缩问题，提出了 FV 的一个二值化版本。

1.1.4 TE

VLAD 和 FV 描述符的构建过程基本上都可以分为两步： 1）嵌入，将低维局部向量映射到高维向量空间。2）聚集，将高维向量集合用 sum-pooling 等方法聚集一个向量。Jégou 等人[48] 对两步分别进行了改进，提出了三角嵌入(TE)和民主聚集(DA)。TE 仅考虑向量的角度而抛弃向量的模长信息，TE 对残差向量进行归一化。TE 计算了特征到每一个码字的残差，而非仅仅计算与最近邻码字的残差。给予主方向太高的权重是冗余的，TE 后会进行一个白化操作。TE 使得不相关的两个向量间的内积接近于 0，即对相似度没有贡献。简单的 sum-pooling 导致各个局部特征在最终得到的全局描述符中的贡献并不一样，DA 使得各个局部描述符对最终的全局描述符的贡献相等。TE 是在嵌入阶段消除不相关向量的交叉作用，而 DA 则是在聚集阶段消除交叉作用。在 Holidays, Oxford 等图像检索数据集上的实验表明，TE 优于 FV-embedding, DA 优于 sum-pooling. TE+DA 性能要优于 FV。TE+DA 后往往还要进行一个 PCA 旋转和归一化处理，以消除 co-occurence 现象[49] ，进一步提高准确率。

1.2 深度特征

1.2.1 全连接层

Babenko 等人[50] 利用预训练的 CNN 来提取图像的特征，并将这些从 CNN 的全连接层提取的特征称为“神经编码（Neural Codes）”。Babenko 等人还进一步研究了“微调(finetune)”对神经编码的影响：在使用一个 Landmark 数据集进行微调的情况下，神经编码在 Oxford building 数据集上的准确率提高了 10 个百分点。Babenko 等人利用 INRIA Holidays、UKB 和 Oxford Buildings 等公共数据集研究了 PCA 对从 CNN 提取的特征的影响，发现神经编码不易受 PCA 降维的影响。

Razavian 等人[51] 研究了 ILSVRC2013 中的基于 CNN 的 Overfeat 网络[52] 提取的特征，表明从 Overfeat 网络提取的特征可以被有效地用于图像分类、场景识别、细粒度识别、属性检测和图像检索等视觉识别领域。Razavian 等人提出的 CNNaug-ss 特征在 Holidays 、UKBench 、Oxford5k 等公共数据集上取得了比普通的 CNN、汉明嵌入（HE）[53] 、VLAD、改进的 Fisher 向量 (IFV)、BOF 等全局特征要好的检索效果。

为了增加 CNN 特征的不变性，Gong 等人[54] 在多个尺度提取图像的 CNN 全连接层特征，并利用它们构成 VLAD 向量，此向量被称为多尺度无序池化（Multi-scale Orderless Pooling， MOP-CNN）。MOP-CNN 对伸缩、平移和旋转都具有一定的不变性。MOP-CNN 在 SUN397 场景识别数据集上的效果要优于 DeCAF,在 Holidays 数据集上的图像检索结果表明，MOP-CNN 的准确率要高于 FV 和 VLAD 等基于 SIFT 的描述符。MOP-CNN 的不足在于，由于采用了滑动窗口法来生成多个尺度的分

© 中国图象图形学报版权所有

片，而每个分片要通过 CNN 一次，增加了生成 MOP-CNN 的时间开销。

Liu 等人[55] 研究表明，FV 并不适合对高维局部向量（如从图像分片提取的 CNN 特征)，并提出了一种新的 FV 编码方法-SCFVC（sparse-coded fisher vector coding）。SCFVC 通过一定的近似，将模型的目标函数转化成了稀疏编码模式。类似于 MOP-CNN, Liu 等人先用滑动窗口法从图片提取大小为 227×227 像素的分片，然后从每一个分片提取一个 FC6 特征（来自 CNN 的 FC6 层）以作为局部特征；最后，利用 SCFVC 对这些局部特征进行编码。不同的是，MOP-CNN 使用了 3 个尺度，而且利用 VLAD 对这些局部 FC 特征进行编码。在场景分类数据集上的实验结果表明，仅使用一个尺度 SCFVC 的性能与使用了 3 个尺度的 MOP-CNN 相当。

同样为了增强描述符的几何不变性，Reddy 等人[56] 提出了对象层深度池化（OLDFP），OLDFP 利用了对象先验知识。OLDFP 利用选择搜索（selective search）[57] 方法来提取包含对象的图像分片，然后再提取各个分片的全连接层特征。“选择搜索”对每幅图片平均生成 2 000 个分片，OLDFP 根据分片包含对象的概率(或得分）来对各分片进行排序，仅选取得分最高的 100 个分片。各个分片的全连接层特征最后通过 max-pooling 来形成一个图片的全局描述符。OLDFP 描述符在各个图像检索公共数据集上取得了优于 TE 和 gVLAD 的性能。OLDFP 在 Holidays 数据集上的 mAP 要高于 MOP_CNN。OLDFP 的不足在于，要利用“选择搜索”来提取图像分片，从中选取的 100 个分片都需要通过 CNN 一次，增加了时间开销。

1.2.2 卷积层

使用卷积层来构造描述符的方法按照其使用的 pooling 方法的不同，主要可以分为 4 类：Max-pooling（最大值池化），sum-poling（和池化），VLAD-pooling（VLAD 池化）和 FV-pooling（FV 池化）。研究者们通常使用 CNN 的最后一个卷积层。使用 pooling 层的方法也算作使用卷积层的方法，因为 pooling 层的特征图（feature map）是对卷积层的特征图使用 max-pooling 得到的。

1）max-pooling。Razavian 等人[58] 对 CNN 生成的图像表述进行了大量的研究。为了增加特征的几何不变性，Razavian 等人提出了多分辨率搜索（MR）。MR 从输入图片不同的位置生成多种尺度的图像分片，因为对象的大小可能不相同，且可能出现在图片的任何位置。MR 将每一个分片输入 CNN，提取最后一个卷积层的特征，进行 max-pooling 便得到单个分片的描述符。两幅图片间的相似度用两幅图片对应的分片描述符集合来计算。为了进一步加强特征的不变性，Razavian 等人提出了 Spatial max-pooling（空间 max-pooling)。Spatial max-pooling 比简单的 max-pooling 保留了更多的空间信息。Razavian 等人从每幅图片提取了 4 种尺度的 30 个分片，这意味着每幅图片要通过 CNN 30 趟，会增大距离计算的开销，Razavian 等人使用 GPU 来进行图片间的距离计算。结合 PCA-白化等后处理技术，MR 在 5 个图像检索数据集上均取得了当时最好的结果。

Tolias 等人[59] 研究了图像检索的“初步搜索”和“再排序”两个过程，提出了卷积区域最大响应（R-MAC）[59] 。R-MAC 利用与 MR[58] 类似的方法来生成多尺度的分片。R-MAC 利用 CNN 的卷积层来生成图像分片的特征，然后用 max-pooling 来形成一个全局描述符。R-MAC 用积分图像来进行近似 max-pooling，以加速对分片的 max-pooling 操作。结合查询扩展（QE），R-MAC 超越了 SPoC[60] 等方法。

细粒度图片检索（FGIR）[61] 源自于细粒度图像识别[62-64]研究的是在由同一种物体形成的数据集中的检索。比如，这种数据集可以是由不同品种的狗形成的，查询时，要找出相似的同一种类的狗。图片之间的差异非常微小。FGIR 通常没有监督数据，所以是一个比通用图像检索要困难得多的问题。通用图像检索只需要找到具有相似的内容（在纹理和形状上）的图片。针对 FGIR 问题，Wei 等人[65] 提

© 中国图象图形学报版权所有

出了选择性卷积描述符聚合（SCDA)[65] 。SCDA 用阈值法来定位图片中的物体，以去掉背景噪声；对此物体区域对应的 CNN 特征进行 average-pooling /max-pooling（实际上是将这两者串联起来，以利用它们的互补性), 便得到了 SCDA。为了提高 SCDA 的区分力，Wei 等人组合 VGGNet 的 pool5 层和 relu5_2 层，形成一个加强版的描述符：SCDA$^+$。为了进一步增强描述符的不变性，Wei 等人串联水平翻转图片的 SCDA$^+$描述符，形成一个新的描述符 SCDA_flip$^+$。SCDA_flip$^+$在 6 个细粒度检索数据集上取得了高于 SPoC 和 CroW[66] 的准确率，与 R-MAC 相当。Wei 等人研究了 SCDA_flip$^+$的压缩问题，发现“SVD+白化”不仅可以对 SCDA_flip$^+$进行降维，还可以进一步提高描述符的准确率，表明 SCDA_flip$^+$中存在冗余。

2）sum-pooling。大部分计算机视觉研究者都是从 CNN 的全连接层提取特征，认为卷积层特征的区分力不够强。Liu 等人[67] 指出，如果适当的使用，卷积层特征可以比全连接层的特征更好。Liu 等人提出了“跨卷积层池化 CCLP) ”[67] 技术，以用来对卷积层特征进行编码。卷积层输出的特征图通常对应了一些具有语义意义的区域，因此 CCLP 利用第(i+1)个卷积层的特征图来为其在第 i 层中对应的 ROI 中的局部特征加权。将第 i 个卷积层中的所有 ROI 中特征的加权和串连起来便得到了最终的图像表示。结合多分辨率方法，CCLP 在场景分类、细粒度分类、物体分类、人类属性分类等任务上取得了优于 MOP-CNN、SCFVC 等使用全连接层的方法。Liu 还指出，简单的符号量化对 CCLP 描述符的区分力影响极其微小，从而可以实现高度压缩。

FV 和 TE 等描述符的构建过程大致分为嵌入和聚集两步。“嵌入”过程将 SIFT 特征嵌入到高维向量空间；“聚集”过程利用 sum-pooling 等方法对高维向量进行汇聚。Babenko 等人[60] 发现，CNN 卷积层特征具有不同于浅层特征（如 SIFT）的特性，利用简单的 sum-pooling 和 PCA-白化对 VGGNet 最后一个卷积层的特征进行处理后形成的描述符在四个公共检索数据集上取得了要高于用全连接层生成的神经编码和 MOP-CNN 的准确率，而不需要费时的高维嵌入过程。此描述符被称为基于和池化的卷积层特征（SPoC）[60] 。由于没有高维嵌入过程，SPoC 的维度不是很高，不需要很多的数据来计算 PCA 矩阵。虽然没有嵌入过程，SPoC 的性能要优于 FV 和 TE 等浅层描述符。

类似于 CCLP[67], Kalantidis 等人提出了跨卷积层加权（CroW）[66] ，以利用 CNN 的最后一个卷积层来生成图像描述符。CroW 在“通道加权（channel weighting）”中引入了稀疏性，因为 Kalantidis 等人发现特征图上的稀疏模式更具有区分力，类似于 BOW 中的稀有特征更具有区分力。此通道加权策略称之为“稀疏敏感的通道加权（SSW)”，类似于 BOW 的 idf 加权，也可以有效地处理“视觉爆发”问题。在低维情况下(128 或 256 维)，CroW 在图像检索公共数据集上取得了优于 Neural Codes，SPoC 和 R-MAC 的结果。

3）VLAD-pooling。Yue 等人[68] 认为全连接层最后面的几层是为了分类的任务而训练的，包含了很多有利于分类的高层语义特征，对局部对象的描述信息不足，不一定适合图像检索；而且训练集图像的尺寸与测试集图像尺寸可能不相同。类似于 MOP-CNN , Yue 等人也利用 VLAD 来对 CNN 特征进行编码，不同的是，MOP-CNN 是在多个尺度对全连接层特征进行 VLAD 编码，而 Yue 等人则是在一个尺度对卷积层进行 VLAD 编码。

为了处理识别图片中的位置，Arandjelovic 等人[69] 提出了一个“端到端学习（end-to-end-learning）”的方法来学习 VLAD 描述符，此描述符被称为 NetVLAD[69] 。Arandjelovic 等人去掉 CNN 后面所有的 FC 层，然后在最后一个卷积层后添加一个 VLAD 层，以码书作为参数，此 VLAD 层可以用 BP 算法来进行训练。此 VLAD 层用的是 Intra-VLAD [41] ，同时结合软分配(soft-assignment) [70] 策略。Arandjelovic 等人用一个卷积层和一个 softmax 层实现了软分配。NetVLAD 在地点识别任务上取得了优于 Intra-VLAD 和全连接层特征的性

© 中国图象图形学报版权所有

能。

4）FV-pooling。类似于 MOP-CNN 从 3 个尺度提取全连接层特征来增强描述符的几何不变性，Yoo 等人[71] 提出了多尺度金字塔池化(MPP)[71] 。Yoo 等人首先对 CNN 进行改造，将后面的全连接层替换成了等价的卷积层，使 CNN 变成一个全卷积层网络，这样输入就可以是任意大小。然后，在此全卷积网络后面添加一个 MPP 层以生成多尺度 FV 描述符。不同于 MOP-CNN 将 3 个尺度的描述符串联，MPP 使用了 average-pooling，对 3 个尺度的 FV 进行了一个平均，形成一个 FV 描述符。尽管 MPP 和 MOP-CNN 都使用了多尺度方法，MPP 与 MOP-CNN 相比，主要的不同点在于：（1）多尺度输入的生成，MOP_CNN 利用滑动窗口法来生成其余两个尺度的图片分片，而 MPP 则利用下采样形成的尺度金字塔；(2)pooling 方法，MOP_CNN 使用的是 VLAD，MPP 使用的是 FV；(3)多尺度融合，MOP_CNN 使用的是串连（将 3 个尺度的特征串连起来），而 MPP 使用的是 average-pooling（对 3 个尺度的 FV 进行平均）。MPP 在场景分类任务上取得了远优于 MOP_CNN 的结果。

1.2.3 softmax 层

自然场景分类是一类具有挑战性的问题，物体在图片的空间布局差异通常很大。基于 SIFT 的 BOW 和基于 CNN 的特征是目前最常用的方法。然而，还有比较特别的方法，这类方法生成的描述符被称为“语义描述符”，此类描述符通常也被称为语义袋 (Bag of Semantics, BoS）[72-73]。BoS 以分类器的概率向量作为特征。尽管此类方法在场景分类任务上取得了优于 BOW 的性能，但是还是不如 FV，因为从图像分片提取的语义特征含有很多噪声，而且 BOW 等编码方法未必直接适用于概率向量。Dixit 等人[74] 提出了 SemanticFV 来解决这些问题。SemanticFV 使用 CNN 分类器来生成语义特征，利用 FV 来对语义特征进行编码。由于 CNN 分类器比传统的 SVM 分类器要准确得多，得到的语义特征更精确。Dixi 等等人认为，由于概率向量空间是非欧的，所以并不适合直接用 FV 对其进行编码。Dixi 等人通过对数变换将概率向量映射到线性空间（欧氏空间）。为了得到足够多的局部语义特征，Dixi 等人从 4 个尺度对图像进行 P×P 的分割，对每个块提取 fc8 层（即 softmax 层)的特征，然后用于生成 SemanticFV。实验表明，SemanticFV 在场景分类数据集上优于使用 fc7 层的 MOP_CNN 和使用 fc6 层的 SCFVC。

1.3 特征的融合

1.3.1 串连

Douze 等人[75] 将属性描述符用于图像检索领域，并与 FV 相结合，取得了当时最好的实验结果。Douze 等人共用了 4 种特征来形成属性描述符：1）SIFT，利用 SIFT 生成 BOF 描述符；2）GIST，描述图像空间布局；3）PHOG(Pyramid of Histograms of Oriented Gradients) [76] ，此描述符描述的是形状；4）Self-similarity 描述符[77] ，此描述符描述的是自相似性，即描述的纹理。Douze 等人将空间金字塔分割与这些描述符相结合以进一步提高区分力，每一个金字塔层对应一个χ^2 -RBF 核，将各个核平均后用于 SVM，每种属性（或类别）对应一个二分类器。将各属性的得分串联起来，标准化并降维后便得到了属性描述符。将属性描述符与 FV 向量串联便得到了融合后的描述符。直接串联是给属性描述符和 FV 赋予了相同的权重，Douze 等人还提出一种加权融合的方法，给 FV 赋予较大的权重，使得准确率提高了 5 个百分点。

1.3.2 核融合

Gehler 等人[78] 提出用多核学习 (MKL)[79] 来将多个核融合成一个单一的模型，以用于处理图像分类问题。每种特征对应一个核，核的融合就是对特征的融合，特征的组合与选择就转化成核的组合与选择问题。MKL 得到的是核的一个最佳的线性组合，也即特征的一个最佳线性组合。MKL 最后只使用了一个 SVM 分类器，Gehler 等人提出可以用线性规划 boosting (LPBoost) [80] ，使用了一种变体 LP-β 来代替 MKL，为 P 种特征训练 P 个 SVM 分类器，每个 SVM 分类器作为一个弱学习机，最

© 中国图象图形学报版权所有

后用各个 SVM 的线性组合来作为一个最终的分类器。此种方式增强了最终分类器的通用性。实验结果表明，LP-β 要优于 MKL 方法，二者都要优于简单的平均核和乘积核，表明 LP-β 和 MKL 都有效的选择具有选择区分力强的特征。

1.3.4 图融合

基于码树[81] 的方法的扩展性很好，但不同查询图片的准确率可能变化很大。Zhang 等人[82] 提出用带权无向图来融合局部特征与全局特征的检索结果，以增强检索的准确率，此方法被称之为基于图的查询融合 (graph-based query specific fusion)，简称为 Graph Fusion（图融合）[82] 。Graph Fusion 对每一种特征的检索结果构建一个 k-互近邻[83] 图，通过对图进行链接分析来对查询结果进行重排序。GraphFusion 以查询结果图片为顶点，以 Jaccard 相似性系数[84] 为顶点间边的权值。结点的连通度反映它与其他图片的相似程度。Zhang 等人提出了两种方法对融合后图中的顶点（图片）进行重排序：1）PageRank[85] 概率向量。根据结点连通度对图片进行排序，利用权值与结点的连通度构建一个随机跳转矩阵，通过迭代算法计算出达到平稳时的各个结点的概率，此概率代表着被访问的概率，也代表着结点的重要程度，按此概率对结点排序并返回。2）加权最大密度子图，返回边的平均权值最大的子图，此子图可以通过贪心算法得到。Zhang 等人以 SIFT 作为局部特征，以哈希化处理的 GIST 和 HSV 作为全局特征，取得了高于码树方法的准确率，而且保留了码树方法的效率与可扩展性。此算法的问题在于要将所有数据库图片的 k-互近邻算出来，而且不能适应数据库的动态变化，最优的 k 值对不同的数据集是不一样的，与查询图片的相关图片数有关。

1.3.5 索引层次的融合

Zhang 等人提出了“语义敏感的协同索引（Semantic aware co-indexing）”[86]，在索引层次融合低层的 SIFT 局部特征和高层的语义属性，以利用它们的互补性增强索引的区分力。很多的融合方法需要在线提取出多种特征，而这增加了时间开销，“语义敏感的协同索引”则是用语义属性离线更线码树索引，仅用 SIFT 进行在线查询。以 denseHOG 和局部二进制模式 (LBP）[87] 来训练 SVM 分类器，以分类器的输出作为语义属性。为了便于进行距离计算，还要用 sigmoid 函数对语义属性进行处理。“语义敏感的协同索引”的生成是离线的，主要包括两步： 1）删除离群图片，对于某个码字对应的倒排索引而言，根据语义属性，删除与其他语义不相似的离群图片。2）插入 K-语义近邻图片（K-semantic nearest neighbor），根据语义属性计算所有数据库图的 K 近邻，并插入到索引项中。查询时，K-语义近邻被用于对 TF-IDF 计算的相似度进行微调。“语义敏感的协同索引”在查询时仅需在线计算 SIFT，语义被融合在索引中。

Liu 等人将 CNN 特征包含进索引中，提出了 DeepIndex[88] 。Liu 等人利用空间金字塔分割来提取 3 个尺度 14 个特征，然后用它们来建立索引。Liu 等人提出了两种形式的 DeepIndex: 1）1-D-DPI，1 维的 DeepIndex, 仅用一个全连接层特征生成的 BOW 索引；2）2-D-DPI，类似于 c-MI[89] 的 2 维 DeepIndex, 用两个全连接层形成的。因为使用了两个 CNN 层，所以包含了两个语义层的信息。对于 2-D-DPI, Liu 等人提出了两个变体：（1）intra-DPI，利用同一个 CNN 的两个全连接层；（2）inter-DPI，两个全连接层分别来自两个 CNN，AlexNet 和 VGGNet。Inter-DPI 的两个全连接层的差异比 intra-DPI 更大，互补性更强。为了提高准确率，Liu 等人计算出图片的全局 CNN 特征存储在一张表格中，作为额外的补充信息，此信息称之 GIS(global image signature)。此 GIS 与 IDF 将一起被用于计算特征间的相似度。与 MA 相结合，2-D-DPI 取得了优于 MOP_CNN 的结果，但要逊色于融合了 SIFT 和 CNN 特征的 DeepEmbedding[90] 。

1.3.6 得分层融合(Score-level fusion)

研究者们通常认为全连接层的特征是最好的，所以只使用 CNN 的全连接层，Li 等人提出了多层无序融合 (Multi-layer Orderless Fusion, MOF）[91]

© 中国图象图形学报版权所有

以融合多个 CNN 层的特征。使用了 CNN-M-128 [92] 的 conv3，conv5 和 fc7 这 3 个层。Conv3 是一个中间层，conv5 是最后一个卷积层，fc7 是一个全连接层。MOF 通过融合 3 个层来同时包含低层的模式和高层的语义。类似于 MOP-CNN, MOF 首先使用滑动窗口法来生成大小为 224×224 像素的图像分片，然后为每一个分片分别提取 conv3,conv5, fc7 层的特征。MOF 使用 max-pooling 对卷积层的特征进行聚合。MOF 为每一层建立一个 BOW 索引，并在索引中分别包含各层特征的 HE 二进制签名。查询时对 3 个层的相似度进行融合(各层相似度的加权和)，各层的相似度计算方式类似于 HE 方法。MOF 在 Holidays, UKB 数据集上取得了与 MOP_CNN 相当的结果。滑动窗口法提取的分片可能含有背景噪声，对描述符的区分力造成了一定程度的影响。

2 国内研究现状

2.1 浅层特征

为了提高描述符的不变性，通常通过在训练集中加入水平翻转后的图片，但是这会使用算法的时空开销加倍。Xie 等人提出了一个对于水平翻转具有不变性的局部描述符 MAX-SIFT[93-94]。MAX-SIFT 是对原 SIFT 和从水平翻转后的图片得到的 SIFT 的进行 max-pooling 得到的。MAX-SIFT 并没有从水平翻转后的图片再提取 SIFT，而是利用原图片的 SIFT 来得到翻转图片的的 SIFT，因为原 SIFT 与翻转图片的 SIFT 间存在一种简单的排列关系。这使得 MAX-SIFT 与 SIFT 的速度相当。在场景分类与细粒度分类等数据集上的实验结果表明，MAX-SIFT 要优于 SIFT。而且也要优于 2014 年提出的“基于狄利克雷分布的直方图特征变换（DHFT)”[95]。虽然 MAX-SIFT 对水平翻转具有不变性，但对于其他严重变形的情形，MAX-SIFT 可能与 SIFT 同样无能为力。

Gao 等人对 TE+DA 编码方法进行了改进，提出了快速 DA（FDA）[96]。 Gao 等人主要从两个方面对 TE+DA 进行了改进：1）提高核矩阵的计算速度。TE 要使用嵌入到高维空间的向量来计算核矩阵，然后利用核矩阵来计算 DA 的权值。但是，Gao 等人认为没有必要将 SIFT 映射到高维向量，对 RootSIFT 进行白化后，whitened-RootSIFT 也可以获得 TE 后的高维向量的属性：相似的特征间的相似度较大，不相似的特征间的相似度较小。利用 whitened-RootSIFT 计算核矩阵极大的减少了算法的时空开销。2）在提取 SIFT 时会引入人工“视觉并发（visual co-occurence）现象”：同一个图像分片被多个 SIFT 表示，这些 SIFT 仅仅是方向不同。Gao 等人提出利用描述符间的空间上下文（如空间位置）来减弱这种现象。Gao 等人利用空间上下文信息提出了一个依赖矩阵，利用依赖矩阵与原来的核矩阵的加权平均来形成一个新的核矩阵。实验结果表明，FDA 比 DA 要快一个量级，而且 FDA 的准确率比 DA 要略高。

2.2 深度特征

2.2.1 全连接层

CNN 的全连接层特征通常有 4 096 维，对其进行 pooling 后得到的维度将更大。Song 等人[97] 提出利用 CNN 来学习一个低维的表示。Song 等人减少 ZFNet[98] 的 FC7 层的滤波器数（一般用 1 024)，以此作为一个瓶颈层，此特征也被称为 DBF(deep bottleneck feature)。Song 等人利用滑动窗口法从输入图提取 224×224 像素的分片，从每一个分片提取一个 DBF, 然后用一种被称之为“二阶池化（second-order pooling）”[99] 的方法来对它们进行聚合，得到的描述符称之为 BoDBF (Bag of DBF)[97]。在 PascalVOC2007[100] 对象分类数据集以及 MIT 场景分类数据集上的结果表明，BoDBF 描述符优于 MOP-CNN[54] 和 SCFV[55]。BoDBF 由于使用的是简单的二阶池化，而且 DBF 特征维度较小，计算量比同样使用全连接层特征的 SCFV、MOP-CNN 等方法要小。

图像分类问题与图像检索有很多相似之处，这两个问题都可以用 BOW 来解决，不同的是，图像检索是“BOW+检索过程”，而分类问题是“BOW+分类器”。Xie 等人表明，图像分类与检索本质上

© 中国图象图形学报版权所有

是相同的，并提出了在线近邻估计（ONE）[101] 这一算法来统一分类与检索问题。ONE 是通过统一分类与检索的相似性计算算法来达到这一目的的。Xie 等人用 3 种方法来对 ONE 算法加速：1)PCA 降维；4 096 维->512 维。2）ANN(approximate nearest neighbor)搜索：用积量化算法编码；3）GPU 加速。ONE 在 3 个场景分类数据集、3 个细粒度分类数据集、2 个图像检索数据集上都取得了当时最先进的结果。尤其是，ONE 在当时最大的场景分类数据集 SUN-397 上使得准确率提升了将近 10 个百分点。ONE 在 Holidays/UKB 数据集上达到了 0.887/3.873。Xie 等人将 SIFT-BOW 与 ONE 相结合，使得检索准确率进一步提升，不过很有限。

MOP-CNN 使用滑动窗口法来生成多个尺度的图像分片，这些分片中有不少含有噪声，而 OLDFP[56] 使用简单的 max-pooling 来对从分片提取的 CNN 特征进行聚合，造成了信息的损失。鉴于此，Bao 等人[102] 使用 VLAD-pooling 来对从各分片提取的全连接层特征进行聚集，由此而形成的描述符称之为基于对象的深度特征聚集 (OADF)[102]。与 OLDFP 相同，OADF 也利用 Selective Search 来生成图像分片，其与 OLDFP 的不同在于对分片使用了 VLAD-pooling，因而可以看做是 OLDFP 的一个改进版。OADF 在 Holidays 和 Oxford 数据集上取得了优于 MOP-CNN 和 OLDFP 的性能。Bao 等人认为可以利用分片之间的关联来进一步提升性能。

2.2.2 卷积层

Gao 等人[103] 研究了用深度特征进行图像识别的系统中各种因子的作用，以得到一个简单、有效和准确率高的图像分类系统。Gao 等人主要研究了 5 种因子：1）层，是使用卷积层还是使用全连接层；2）标准化；3）FV 的 GMM 的 Gaussian 分量数(K)；4）空间信息；5）多尺度。基于对这些影响因子的研究，提出了深度空间金字搭(DSP）[103] 描述符。DSP 对卷积层特征图（而非对输入图像）使用空间金字塔(SP)分割以捕获空间信息，然后使用 IFV 对分割后的每一个块进行编码，将各个块的 IFV 描述符串联起来后便得到 DSP 描述符。因为是在特征图上使用 SP 分割，而不是在原输入图像上进行，所以只需前向通过 CNN 一次，节省了时间开销。MPP[71] 对原输入图像构建多分辨率金字塔，MOP-CNN 则是对输入图像使用滑动窗口法来获取多个尺度的信息。此外，Gao 等人提出了一个新的特征标准化方法：2-范数矩阵标准化，使用图片所有卷积层特征形成的矩阵的谱范数来对特征进行标准化。2-范数矩阵标准化使用了来自整幅图像的信息，可以捕获一些全局信息，对光照和尺度变化等更具有较强的抵抗力。Gao 等人发现，当 FV 的 K 在 1~4 之间时（基于 SIFT 的 FV 使用的 K 值通常位于 64~ 256），DSP 即可取得最优的结果。如此小的 K 值将极大的减小 DSP 的维度。极小的 K 值之所以有效，Gao 等人认为是由于从卷积层提的局部特征太少了(100 个左右)，不足以用于准确估计较大的 GMM 模型。为了进一步捕获多个尺度的信息，Gao 等人提出了 DSP 的一个多尺度版本 -Multi-scale DSP(Ms-DSP)。Ms-DSP 对 5 个尺度的输入图片提取 DSP，然后取平均。DSP 在对象识别、场景识别、动作识别等数据集上都取得较好的结果。

2.3 特征融合

2.3.1 串连

为了同时利用局部特征与全局特征，以增强特征的区分力与抗噪性能，Sun 等人提出了 OR[104]。Sun 等人首先利用 BING(二值化梯度范数)[105] 来获取图像分片；从每一个分片提取一个 VLAD 描述符(用 SIFT 生成)和一个 CNN 描述符(全连接特征），PCA-whitening 后串连起来即为 OR 描述符。Sun 等人进一步利用积量化和倒排索引来加速检索，以适应大规模数据集。

近年来 CNN 在各个领域都取得当前最先进的结果，那么我们是否可以抛弃 SIFT 直接使用 CNN 特征呢？Yan 等人认为，SIFT 与 CNN 是互补的关系，并提出了 CCS(complementary CNN and SIFT)[106] 描述符来融合 SIFT 与 CNN 特征。CCS 是一种多层表示，融合了多个层次的信息：1）场景层。场景层代表的是高层的语义信息，提取 GoogLeNet 的

© 中国图象图形学报版权所有

pool5 层作为此层的表示。2）对象层。利用 EdgeBox[107] 提取图像分片，选取得分最高的前 100 个分片，从每一个分片提取 pool5 层的特征，然后用于生成 VLAD。3）点层。利用 SIFT 生成 VLAD。融合 SIFT 可以有效的提高描述符的几何不变性。CCS 利用 PCA 将 VLAD 降到 1 024 维，然后将 3 个层次的描述符串连起来并进行归一化，进行 PCA-whitening 处理，再归一化后即为最终的 CCS 描述符。CCS 在 Oxford 数据集、Paris 数据集和 UKB 数据集上取得了优于 SPoC, MOP_CNN, OLDFP 的准确率，证明融合 SIFT 与 CNN 的有效性。不过，CCS 的 VLAD 的码书很大，取的是 500（一般的情况下只取 64~256），这么大的码书会导致 VLAD 的维度很高，PCA 矩阵的计算将比较困难。

为了利用多种互补的特征，Ge 等人[108] 提出利用稀疏编码[109] 来对不同的特征进行编码，然后串连起来以达到融合的目的，此描述符在此处称之为“稀疏编码的特征（SCF)”[108] 。对于每一种特征，Ge 等人先利用稀疏编码方法来生成特征的稀疏编码，然后利用 max-pooling 对稀疏编码进行聚合。Ge 等人研究了特征的检测子与描述子的组合问题，提出利用 Harris-DAISY[110] 和 LOG-SIFT 两种局部特征描述符。另外，Ge 等人还提出了一种新局部颜色描述符 micro。利用“稀疏编码+max-pooling”对所有分片的 micro 特征进行编码，就形成了一个新的颜色描述符-Sparse-coded micro feature (SCMF)。micro 特征利用了图片的自相似性。将 Harris-DAISY, LOG-SIFT 和 micro 3 个特征的稀疏编码描述符串连起来就是最终的融合描述符。Ge 等人还研究了用 PCA 和积量化对此描述符进行压缩以用于大规模图像检索的情形。实验表明，此描述符要优于 VLAD, FV 以及颜色袋（BOC）[111] 。虽然此稀疏编码特征在 UKB 上超越了当时的其他方法，但是，在 Holidays 数据集上要逊色于 LBOC(local BoC)[111] 和 HE。

2.3.2 核融合

简单的串连会增加特征的维度，Yeh 等人提出用 MKL 来融合来自不同域的特征，此法称之为 GL-MKL (group lasso muti-kernel learning)[112] 。GL-MKL 使每种特征对应多个核，以每种特征作为一个组。GL-MKL 混合使用 ℓ_1-范数约束和 ℓ_2-范数约束（称为 $\ell_{1,2}$-norm 约束)，以作为一个组 lasso 约束子(group lasso regularizer)。GL-MKL 使用 MKL 来学习每个组中核的权值。组 lasso 约束子增强了组间的稀疏性，但组内却不用是稀疏的。组间稀疏性使得仅有少数区分力强的特征被使用，所以 GL-MKL 也是一种特征选择方法。Yeh 等人将 GL-MKL 用于视频物体分类和图片分类。在处理视频物体分类问题时，使用了 MFCC (梅尔频率倒谱系数)音频特征，SIFT 特征，HOG 特征，Gabor 滤波器[113] 和 EDH(边缘方向直方图)[114] 。实验结果表明，GL-MKL 要优于 LP-β[114] 。本质上，GL-MKL 可以看成是 LPBoosting 的进一步推广，引入了组稀疏性的概念。

2.3.3 图融合

Liu 等人[115-116] 认为 GraphFusion 方法易受离群图片(outliers)的影响，因为：1）特征。并不是所有特征都是有效的，无效特征会引入离群图片；2）K 近邻数。GraphFusion 使用的是 K-互近邻，K 值理论上应当与查询图片的真实相关图片(groundtruth)数相等，但每张查询图片的相关图片数是不一样的，如果 K 大于相关图片数，就会引入离群图片。鉴于此，Liu 等人提出了一个更不易受离群图片影响的方法——ImageGraph[115-116] 。ImageGraph 是 GraphFusion 方法的改进版，与 GraphFusion 方法有如下不同：（1）图。GraphFusion 用的是 K-互近邻图，而 ImageGraph 用的是一个单向 K-近邻图（仅含出边，指向 K-近邻）。K-近邻图的结点数比 K-互近邻图的结点数更多，可以提高检索的查全率。（2）相关性度量。Liu 等人提出一个被称为 Rank Distance 的方法来度量两幅图片的相关程度。Rank Distance 利用了两幅图片的排序，不易受离群图片的影响。（3）相似性度量。GraphFusion 用的是 Jaccard 相似性，ImageGraph 的相似性度量方法称之为“贝叶斯相似性”，是基于 Rank Distance 方法求的概率模型。（3）排序方法。

© 中国图象图形学报版权所有

GraphFusion 的排序方法可能会导致不相关的图片间有很多边。ImageGraph 的方法称为“Local Ranking”。Local Ranking 旨在寻找一个最大加权子图，是一个局部最优的方法，而非全局最优，以避免被紧密相连的离群图片影响。ImageGraph 用到了 SIFT，GIST,HSV 和 CNN 特征，并在 Holidays 和 UKB 数据集上取得了优于 GraphFusion 和“查询自适应晚期融合”[117] 的结果。

2.3.4 索引层融合

从 IMI (inverted multi-index)[89] 得到启发，Zheng 等人提出了耦合多维索引 (coupled Multi-Index, c-MI) [9]，对 SIFT 和 CN(颜色名)[118] 颜色特征在索引层次进行了融合。c-MI 是一个 2 维索引，以 SIFT 和 CN 分别作为索引的 1 维。SIFT 和 CN 分别对应一个码书，它们的每一个码字组合对应一个倒排列表。查询时，取出码字组合对应的倒排列表，tf-idf 及 CN 的二进制签名计算相似度。Zheng 等人将 c-MI 的策略总结为“装箱(packing)”和“填补(padding)”:“装箱”是指以 SIFT 和 CN 分别作为索引的 1 维；”填补”则是指使用一些别的策略来进一步提高检索的准确率与查全率。具体“填补”的内容有：1）MA(Multiple Assignment，多分配)。取多个近邻的倒排列表以提高查全率。2）SIFT 的 HE 二进制签名；3）burstiness 加权[42]；消除“视觉爆发现象”。4）Graph Fusion。可以利用 Graph Fusion 将 c-MI 的结果与 HSV 的结果进行融合。c-MI 在 Holidays、UKB 等图像检索公共数据集上取得了当时最好的结果。c-MI 不仅时空开销较小，而且还可以进一步与其他的特征融合，不过索引的维度越高，倒排列表将会越稀疏，要提高查全率与准确率就要访问更多的倒排列表。

为了有效的同时利用 SIFT 和 CNN 特征，Zhou 等人提出了“协同索引嵌入（Collaborative Index Embedding，CIE）[119]。CIE 利用索引矩阵对 SIFT 和 CNN 的两个特征空间的图片近邻结构进行相互迭代校正，使两个特征空间的近邻结构尽可能相似。由于两个特征空间的近邻结构接近，所以最后在查询时，只需 CNN 特征即可。CIE 将 CNN(AlexNet) 的两个全连接层特征串连起来以作为 CNN 特征，并通过域值化对其进行了稀疏化处理，以适应于索引，减少索引开销，加快查询速度。CIE 在 Holidays 与 UKB 数据集上取得了与“查询自适应晚期融合”[117] 以及 ONE[101] 相当的准确率。

2.3.5 得分层融合(Score-level fusion)

在利用多个特征进行检索时，对于给定的特征，并不知道哪些特征是有效的，哪些特征是无效的，所以应当开发一种自适应查询的方法。Zheng 等人提出了一种得分层（score-level）多特征融合方法-查询自适应晚期融合（query adaptive late fusion）[117]。Zheng 等人的动机在于他们发现：对于一个好的特征而言，其排序后的得分曲线应该是 L 型的（先快速下降，然后趋于平稳），而不好的特征的得分曲线是逐渐下降的。“查询自适应晚期融合”主要有两个特点：（1）以查询自适应的方式估计特征的有效性。各特征的权值是不固定的，不易受无效特征的不良影响。（2）特征的有效性是利用无关数据集在线估计的，不依赖数据库本身，可以适应大规模数据库的动态变化。Zheng 等人利用不相关的数据集近似数据库图片上的 score 曲线的尾部，用特征的 score 曲线减去此尾部以突出 top-k 图片的作用，此 score 曲线与坐标轴围成的面积的倒数便反映了特征的好坏（或有效性）。以此面积的倒数作为对应特征的权值以计算相似度。Zheng 等人共利用了 5 种特征：SIFT、HSV、CaffeNet[120] 全连接层特征、GIST 和随机特征。GIST 与随机特征主要作为无效特征，以测试算法的健壮性。实验表明，“查询自适应晚期融合”要优于图融合和索引层次融合的协同索引[86]。

BOW 仅使用 SIFT 特征来进行匹配，而 SIFT 特征仅代表了局部的信息，忽视了其他的信息，而且被量化到同一个视觉单词的特征即认为是匹配的，这会导致大量的虚假匹配(false match)。Zheng 等人[90] 认为，一对关键点要成为真实匹配(true match)，需要在“局部/区域/全局”3 个层次上匹配。为达到这一目的，Zheng 等人提出了一个 DeepEmbedding[90] 框架，利用 3 个层次的信息来为匹配过程建立一个概率模型，此模型就是 CNN

© 中国图象图形学报版权所有

特征与 SIFT 的融合模型。Zheng 等人利用空间金字塔来对图像划分，总共分为 3 个尺度（1×1, 4×4, 8×8），除 global 尺度外，剩下的两个尺度用于区域层。全局层与区域层均由 Decaf 网络的全连接层特征来描述，而局部层则用 SIFT 来描述，区域信息与全局信息被称之为 SIFT 特征点的上下文环境。Zheng 等人还提出一个 DeepIndexing 的索引结构，仅在索引中存放 SIFT 的 HE 签名与区域、全局特征的指针，而 regional 与 global 特征的 LSH（局部敏感哈希）[122] 签名则统一放在外部的表格中。与 MA 等策略相结合，DeepEmbedding 在 Holidays、UKB 等数据集上取得了优于在索引层次融合 SIFT 与颜色特征的 c-MI 的性能。

3 挑战及趋势

3.1 讨论

卷积层特征与 SIFT 相比，有如下特点：1）卷积层特征类似于密集 SIFT 特征（通过网格式的密集采样得到）。卷积层特征与 SIFT 一样是局部特征，对应了图片的某个区域（可以将 CNN 特征图上每一个点反向映射回图片），是一种局部特征。2）卷积层特征是通过学习得到的，SIFT 是手工类型。CNN 的卷积层参数是可以针对不同的数据集通过迭代训练调优的，而且可通过简单的修改进一步改进（如增加深度、宽度等）而 SIFT 的参数是通过预先的精密设计固定的。3）卷积层特征具有层次性。不同的卷积层具有不同的语义层次[98]，如浅层的特征图通常是一些边/角等，而中层则是物体的一部分,高层则通常是一个完整的物体。选用不同的层将可能达到完全不同的效果，该如何选择一个最优的层则到目前为止还没有一个最优的方法，通常通过测试多层的效果来达到。SIFT 在不使用 SP 的情况下不具有层次性，描述的是边/角等比较低层次的特征，这也是为什么 CCS[106] 将 SIFT 与 CNN 融合会有效果的原因之一。4）CNN 卷积层特征维度比 SIFT/SURF 等浅层特征要大得多，而且计算量大，需要 GPU 辅助才能达到实时的效果，而且因为要存储很多卷积层特征图的原因，空间开销也要大得多。对于 PC 机而言，这不是什么大问题，然而未来的 AI 将可能无处不在，CNN 在移动平台上的使用将成为一个具有挑战性的问题。随着类脑计算[125]如火如荼的展开，各种神经处理专用芯片（如中国科学院陈云霁等人研发的 DaDianNao[126]，Google 最近研发的 TPU[127] 等）不断涌现，此问题或者也将不是问题。

3.2 挑战

3.2.1 SIFT

SIFT 存在两个问题： 1）视觉爆发现象[42]：大量的特征被分配到少量的视觉单词。对于具有自相似性的纹理图片而言，不少 SIFT 从 2-范数距离意义上来讲是很相似的，这会导致匹配时的假正（false positive）现象。2）视觉并发现象[49]：码字的出现并不是相互独立的。这两个问题通常在编码阶段被处理，如 VLAD 为了处理“视觉爆发现象”采用的内部标准化策略与 IFV 采用的幂律标准化策略。“视觉并发现象”通常采用 PCA-whitening[121]来处理。虽然都很有效，不过并没有谁找到一种最优的方法来处理这两个问题，而且这两种现象和数据集类型有很大的关系。

编码方法通常分为两个步骤： 1）嵌入。将低维局部特征嵌入（或映射）到高维空间。此过程通常是为了采集局部特征分布的统计信息。如 BOF 只含有 0 阶统计信息（频数），VLAD 含有 1 阶统计信息（均值），FV 含有 1 阶（均值）和 2 阶（方差）的统计信息。不过，目前这些编码方法基本上都是基于局部特征的分布属性的方法，而且这些分布属性基本上是采用 k-means、GMM 等非监督聚类方法得到的。k-means 等方法是基于 2-范数距离的，由于“维度灾难(curse-of-dimensionality)”[122] 的存在，高维向量在欧氏空间具有高度的相似性，不易区分，因而 2-范数距离并非最优的。而且高维向量中通常含有大量的信息冗余。将高难空间映射到低维空间来处理，从一定程度上能缓解此问题，然而降维会造成部分有用信息的损失。2）汇聚。局部特征的数目通常非常多，而且会随图片的大小与特

© 中国图象图形学报版权所有

征的类型而变化，sum-pooling（通常是加权平均）在编码方法中被广泛的使用，以消除特征数目的影响；而 max-pooling 则在对 CNN 的卷积层汇聚时用的最多，因为 max-pooling 对微小的变化具有一定的不变性。不同的局部特征(SIFT/CNN 卷积层特征）来自不同的位置，描述能力（或区分力）也不一样，因而对特征加权是现在经常使用并且在将来还会继续被广泛使用的策略。

SIFT 虽然具有很强的几何不变性，但是其本身缺乏几何信息（如尺度、方向、位置），所以通常通过增强几何信息来增强区分力。目前主要通过 3 种策略来达到这一目的： 1）SIFT 层。扩展 SIFT 描述符，将几何信息串连在 SIFT 后面。2）编码层。如 gVLAD[43]，利用 SIFT 的主方向来生成包含方向信息的 VLAD。3)索引层(或 score 层)。如 HE（Hamming embedding)[53] 使用尺度与角度信息来校正相似度，HE 将其称之为“弱几何一致性(WGC)”。有“弱”就有强，强几何一致性通常通过仿射匹配来达到，称之为“空间验证（SV)”[1] [46]。SV 同时考虑了位置、尺度、方向 3 个因子，利用它们来建立仿射模型，利用 RANSAC(random sample consensus) [123] 来迭代校正模型，最后用仿射模型来验证图片的几何一致性。不过，RANSAC 只能用在两个集合之间，只适合局部特征，并不适合全局特征，因为一幅图片只能生成一个全局特征，所以 SV 通常用于 SIFT-BOW 模型。所以，全局特征目前缺乏强几何一致性的验证方法，一般只能通过编码来包含少量的几何信息，或者通过空间金字塔[124] 来达到。但空间金字塔会使得描述符的维度成倍增加，在大规模的情况下，会增加不少的时空开销。

3.2.2 CNN 特征

尽管 CNN 特征目前在图像检索领域被广泛使用，但与 SIFT 相比，CNN 特征在图像检索方面还存在如下不足：1）通用性。SIFT 可以被用于任意数据集，不需要考虑数据集的分布。而用 ImageNet 预训练的 CNN 特征则在通用性方面要差一些，目标数据集与 ImageNet 的差异越大，图像检索的性能就会越差。用与目标数据集相近的数据集重新训练 CNN 几乎是不可能的，因为这种带标签的大数据集一般是没有的，而小数据集会导致过拟合问题。故而此问题通常通过用与目标数据集相近的小数据集微调预训练的 CNN 来解决，然而哪怕是收集这种小数据集也很费事，因为还要人工标注。2）几何不变性。CNN 特征与 SIFT 这种局特征相比，在尺度、旋转、平移及光照变化等方面的不变性要差得多。虽然 MOP_CNN[54] 通过串联 3 个尺度的 VLAD 增强了尺度不变性，MOP_CNN 对于旋转、平移等因素的不变性却没有得到处理。也可以在源头解决 CNN 特征的不变性问题，那就是“数据增强”：将训练图片经过旋转、伸缩、平移等处理后的图片也加入训练集。不过这会使得训练的时空开销成倍增加。3）特征数。从一幅图片一般可以提取几千个 SIFT，即使是小图片，也可以通过密集采样得到数目众多的SIFT。而从一幅图片提取的CNN特征则很少：1 个全连接层特征或几百个卷积层特征。一般通过生成很多图像分片来解决此问题，但每个分片要通过 CNN 一趟，会使 CNN 特征提取的开销增加。

3.3 趋势

针对 CNN 特征的不足，未来的特征及特征融合方法的趋势可能如下：1)CNN 架构。目前的 CNN 架构几乎都是用于图像分类问题的，然而图像分类问题与图像检索问题有很大的不同，图像检索是一个更细粒度的问题，更观注图像包含的局部视觉模式，检索算法对这种模式的区分力是一个图像检索问题的一个决定因素。增加 CNN 对图片模式的区分力，找到一种更适合图像检索的 CNN 架构是一个值得研究的问题。2）几何不变性。CNN 特征缺乏几何不变性，但目前研究者们对此问题进行处理的人比较少。增强 CNN 特征对尺度、旋转、平移及光照变化等各种因素的不变性，毫无疑问，将显著提升检索算法的准确率，然而时空开销的增加将不可避免。3）特征融合。CNN 特征缺乏不变性，在不改变其本身的情况下，可以通过融合不变性强的特征来解决此问题；再者，通过融合互补的多种

© 中国图象图形学报版权所有

特征可以有效的增加描述符的区分力。然而，特征的有效性及融合方法依然需要研究。“查询自适应晚期融合”[119]通过研究 Score 曲线的形状来判定特征的有效性，然而此种方法的代价较大，能否找到一种在”早期“（比如编码阶段）就能判定特征是否有效的方法呢？这个问题值得思考。未来的图像检索将无处不在，各种平台在存储与计算能力方面的差异都将为本领域带来挑战，如何权衡好速度、空间开销、准确率等方面，依然是图像检索领域将要面对的问题。

志谢

本报告的撰写得到中国计算学会多媒体专业委员会、华中科技大学、浙江大学和清华大学相关研究团队研究人员的大力支持，特此致谢。

参考文献

[1] Lowe D G. Object recognition from local scale invariant feature [J]. Proceedings of the IEEE International Conference on Computer Vision. 1999, 2:1150 - 1157.

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

[3] 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.

[4] Sivic J, Zisserman A. Video google: A text retrieval approach to object matching in videos[C]. ICCV 2003, 2(1470): 1470-1477.

[5] Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE, 2007: 1-8.

[6] Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation[C]. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010: 3304-3311.

[7] Douze M, Ramisa A, Schmid C. Combining attributes and fisher vectors for efficient image retrieval[C]. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2011: 745-752.

[8] Zhang S, Yang M, Cour T, et al. Query specific fusion for image retrieval[J]. Computer Vision–ECCV 2012, 2012: 660-673.

[9] Zheng L, Wang S, Liu Z, et al. Packing and padding: Coupled multi-index for accurate image retrieval[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1939-1946.

[10] LeCun Y, Boser B E, Denker J S, et al. Handwritten digit recognition with a back-propagation network[C]. Advances in neural information processing systems. 1990: 396-404.

[11] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4): 541-551.

[12] 吴飞，朱文武，于俊清,等. 多媒体技术研究:2014——深度学习与媒体计算[J]. 中国图象图形学报, 2015, 20(11):1423-1433.

[13] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 248-255.

[14] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems. 2012: 1097-1105.

[15] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]. Proceedings of the 27th international conference on machine learning (ICML-10). 2010: 807-814.

[16] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4): 212-223.

[17] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[19] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[20] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[21] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[22] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[23] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 221-231.

[24] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]. Advances in neural information processing systems. 2014: 568-576.

[25] Hannun A, Case C, Casper J, et al. Deep speech: Scaling up end-to-end speech recognition[J]. arXiv preprint arXiv:1412.5567, 2014.

[26] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep speech 2: End-to-end speech recognition in english and mandarin[C]. International Conference on Machine Learning. 2016: 173-182.

[27] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]. Advances in neural information processing systems. 2014: 3104-3112.

[28] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

© 中国图象图形学报版权所有

[29] Donahue J, Jia Y, Vinyals O, et al. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition[C]. ICML. 2014, 32: 647-655.

[30] Sharif Razavian A, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: an astounding baseline for recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014: 806-813.

[31] Azizpour H, Razavian A S, Sullivan J, et al. Factors of transferability for a generic convnet representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(9): 1790-1802.

[32] Zhao R, Grosky W I. Narrowing the semantic gap-improved text-based web document retrieval using visual features[J]. IEEE transactions on multimedia, 2002, 4(2): 189-200.

[33] Hare J S, Lewis P H, Enser P G B, et al. Mind the Gap: Another look at the problem of the semantic gap in image retrieval[C].Electronic imaging 2006. International Society for Optics and Photonics, 2006: 607309-607309-12.

[34] Ke Y, Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors[C]. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004, 2: II-506-II-513 Vol. 2.

[35] Mikolajczyk K, Schmid C. A performance evaluation of local descriptors[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, .2005, 27(10): 1615-1630.

[36] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[C].Computer vision–ECCV 2006. Springer Berlin Heidelberg, 2006, 404-417.

[37] Arandjelovic R, Zisserman A. Three things everyone should know to improve object retrieval[C].2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, 2911-2918.

[38] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.

[39] Jégou H, Douze M, Schmid C. Packing bag-of-features[C]. 2009 IEEE 12th International Conference on Computer Vision. 2009, 2357-2364.

[40] Jegou 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.

[41] Arandjelovic R, Zisserman A. All about VLAD[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 1578-1585.

[42] Jégou H, Douze M, Schmid C. On the burstiness of visual elements[C]. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 1169-1176.

[43] Wang Z, Di W, Bhardwaj A, et al. Geometric VLAD for large scale image search[J]. arXiv preprint arXiv:1403.3829, 2014.

[44] Sánchez 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.

[45] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification[J]. Computer Vision–ECCV 2010, 2010: 143-156.

[46] Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching[C].Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007: 1-8.

[47] Perronnin F, Liu Y, Sánchez J, et al. Large-scale image retrieval with compressed fisher vectors[C]. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), .IEEE, 2010: 3384-3391.

[48] Jégou H, Zisserman A. Triangulation embedding and democratic aggregation for image search[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 3310-3317.

[49] Chum O, Matas J. Unsupervised discovery of co-occurrence in sparse high dimensional data[C]. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2010: 3416-3423.

[50] Babenko A, Slesarev A, Chigorin A, et al. Neural codes for image retrieval[C]. European conference on computer vision. Springer International Publishing, 2014: 584-599.

[51] Razavian A S, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: an astounding baseline for recognition[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2014, 512-519.

[52] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv:1312.6229, 2013.

[53] Jegou H, Douze M, Schmid C. Hamming embedding and weak geometric consistency for large scale image search[C]. Computer Vision–ECCV 2008. Springer Berlin Heidelberg, 2008: 304-317.

[54] Gong Y, Wang L, Guo R, et al. Multi-scale orderless pooling of deep convolutional activation features[C].Computer Vision–ECCV 2014. Springer International Publishing, 2014, 392-407.

[55] Liu L, Shen C, Wang L, et al. Encoding high dimensional local features by sparse coding based fisher vectors[C]. Advances in Neural Information Processing Systems. 2014: 1143-1151.

[56] Reddy Mopuri K, Venkatesh Babu R. Object level deep feature pooling for compact image representation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015: 62-70.

[57] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154-171.

[58] Razavian A S, Sullivan J, Carlsson S, et al. Visual instance retrieval with deep convolutional networks[J]. arXiv preprint arXiv:1412.6574, 2014.

[59] Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations[J]. arXiv preprint arXiv:1511.05879, 2015.

[60] Babenko A, Lempitsky V. Aggregating local deep features for image retrieval[C]. Proceedings of the IEEE international conference on computer vision. 2015: 1269-1277.

[61] Xie L, Wang J, Zhang B, et al. Fine-grained image search[J]. IEEE Transactions on Multimedia, 2015, 17(5): 636-647.

[62] Lin T Y, RoyChowdhury A, Maji S. Bilinear cnn models for fine-grained visual recognition[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 1449-1457.

[63] Krause J, Jin H, Yang J, et al. Fine-grained recognition without part annotations[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5546-5555.

[64] Zhang Y, Wei X S, Wu J, et al. Weakly supervised fine-grained categorization with part-based image representation[J]. IEEE Transactions on Image Processing, 2016, 25(4): 1713-1725.

© 中国图象图形学报版权所有

[65] Wei X S, Luo J H, Wu J, et al. Selective convolutional descriptor aggregation for fine-grained image retrieval[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2868-2881.

[66] Kalantidis Y, Mellina C, Osindero S. Cross-dimensional weighting for aggregated deep convolutional features[C]. European Conference on Computer Vision. Springer International Publishing, 2016: 685-701.

[67] Liu L, Shen C, van den Hengel A. The treasure beneath convolutional layers: Cross- convolutional-layer pooling for image classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4749-4757.

[68] Yue-Hei Ng J, Yang F, Davis L S. Exploiting local features from deep networks for image retrieval[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015: 53-61.

[69] Arandjelovic R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5297-5307.

[70] 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. CVPR 2008. IEEE, 2008: 1-8.

[71] Yoo D, Park S, Lee J Y, et al. Fisher kernel for deep neural activations[J]. arXiv preprint arXiv:1412.1628, 2014.

[72] Torresani L, Szummer M, Fitzgibbon A. Efficient object category recognition using classemes[J]. Computer Vision–ECCV 2010, 2010: 776-789.

[73] Kwitt R, Vasconcelos N, Rasiwasia N. Scene recognition on the semantic manifold[J]. Computer Vision–ECCV 2012, 2012: 359-372.

[74] Dixit M, Chen S, Gao D, et al. Scene classification with semantic fisher vectors[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2974-2983.

[75] Douze M, Ramisa A, Schmid C. Combining attributes and fisher vectors for efficient image retrieval[C]. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2011: 745-752.

[76] Bosch A, Zisserman A, Munoz X. Representing shape with a spatial pyramid kernel[C]. Proceedings of the 6th ACM international conference on Image and video retrieval. ACM, 2007: 401-408.

[77] Shechtman E, Irani M. Matching local self-similarities across images and videos[C]. IEEE Conference on. Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE, 2007: 1-8.

[78] Gehler P, Nowozin S. On feature combination for multiclass object classification[C]. 2009 IEEE 12th International Conference on Computer Vision . IEEE, 2009: 221-228.

[79] Lanckriet G R G, Christianini N, Bartlett P L, et al. Learning the Kernel Matrix with Semi-Definite Programming[C]. Nineteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2002:323-330.

[80] Demiriz A, Bennett K P, Shawe-Taylor J. Linear Programming Boosting via Column Generation[J]. Machine Learning, 2002, 46(1-3):225-254.

[81] Nister D, Stewenius H. Scalable recognition with a vocabulary tree[C]. 2006 IEEE computer society conference onComputer vision and pattern recognition. IEEE, 2006, 2: 2161-2168.

[82] Zhang S, Yang M, Cour T, et al. Query specific fusion for image retrieval[J]. Computer Vision–ECCV 2012, 2012: 660-673.

[83] Qin D, Gammeter S, Bossard L, et al. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors[C].2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2011: 777-784.

[84] Jaccard P. The distribution of the flora in the alpine zone[J]. New phytologist, 1912, 11(2): 37-50.

[85] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab, 1999.

[86] Zhang S, Yang M, Wang X, et al. Semantic-aware co-indexing for image retrieval[C]. Proceedings of the IEEE International Conference on Computer Vision. 2013: 1673-1680.

[87] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 24(7): 971-987.

[88] Liu Y, Guo Y, Wu S, et al. Deepindex for accurate and efficient image retrieval[C]. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, 2015: 43-50.

[89] Babenko A, Lempitsky V. The inverted multi-index[C]. 2012 IEEE Conference onComputer Vision and Pattern Recognition (CVPR). IEEE, 2012: 3069-3076.

[90] Zheng L, Wang S, He F, et al. Seeing the big picture: Deep embedding with contextual evidences[J]. arXiv preprint arXiv:1406.0132, 2014.

[91] Li Y, Kong X, Zheng L, et al. Exploiting Hierarchical Activations of Neural Network for Image Retrieval[C]. Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016: 132-136.

[92] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets[J]. arXiv preprint arXiv:1405.3531, 2014.

[93] Xie L, Tian Q, Zhang B. Max-SIFT: Flipping invariant descriptors for Web logo search[C]. 2014 IEEE International Conference on Image Processing (ICIP) . IEEE, 2014: 5716-5720.

[94] Xie L, Tian Q, Wang J, et al. Image classification with Max-SIFT descriptors[C].IEEE International Conference on Acoustics, Speech and Signal Processing. 2015.

[95] Kobayashi T. Dirichlet-based histogram feature transform for image classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 3278-3285.

[96] Gao Z, Xue J, Zhou W, et al. Fast democratic aggregation and query fusion for image search[C]. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, 2015: 35-42.

[97] Song Y, McloughLin I, Dai L. Deep Bottleneck Feature for Image Classification[C]. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, 2015: 491-494.

[98] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]. European conference on computer vision. Springer International Publishing, 2014: 818-833.

[99] Carreira J, Caseiro R, Batista J, et al. Semantic segmentation with second-order pooling[J]. Computer Vision–ECCV 2012, 2012: 430-443.

[100] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88(2): 303-338.

© 中国图象图形学报版权所有

[101] Xie L, Hong R, Zhang B, et al. Image classification and retrieval are one[C].Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. 2015: 3-10.

[102] Bao Y, Li H. Object-Based Aggregation of Deep Features for Image Retrieval[C]. International Conference on Multimedia Modeling. Springer, Cham, 2017: 478-489.

[103] Gao B B, Wei X S, Wu J, et al. Deep spatial pyramid: The devil is once again in the details[J]. arXiv preprint arXiv:1504.05277, 2015.

[104] Sun S, Zhou W, Tian Q, et al. Scalable object retrieval with compact image representation from generic object regions[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2016, 12(2): 29.

[105] Cheng M M, Zhang Z, Lin W Y, et al. BING: Binarized normed gradients for objectness estimation at 300fps[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 3286-3293.

[106] Yan K, Wang Y, Liang D, et al. CNN vs. SIFT for Image Retrieval: Alternative or Complementary?[C]. ACM on Multimedia Conference. ACM, 2016:407-411.

[107] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges[C]. European Conference on Computer Vision. Springer International Publishing, 2014: 391-405.

[108] Ge T, Ke Q, Sun J. Sparse-Coded Features for Image Retrieval[C]. BMVC. 2013.

[109] Yang J, Yu K, Gong Y, et al. Linear spatial pyramid matching using sparse coding for image classification[C]. 2009. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 1794-1801.

[110] Winder S, Hua G, Brown M. Picking the best daisy[C]. 2009. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 178-185.

[111] Wengert C, Douze M, Jégou H. Bag-of-colors for improved image search[C]. Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011: 1437-1440.

[112] Yeh Y R, Lin T C, Chung Y Y, et al. A Novel Multiple Kernel Learning Framework for Heterogeneous Feature Fusion and Variable Selection[J]. IEEE Transactions on Multimedia, 2012, 14(3):563-574.

[113] Manjunath B S, Ma W Y. Texture Features for Browsing and Retrieval of Image Data.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1996, 18(8):837-842.

[114] Yanagawa A, Hsu W, Chang S F. Brief descriptions of visual features for baseline TRECVID concept detectors[J]. 2006.

[115] Liu Z, Wang S, Zheng L, et al. Visual reranking with improved image graph[C].Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014: 6889-3893.

[116] Liu Z, Wang S, Zheng L, et al. Robust ImageGraph: Rank-Level Feature Fusion for Image Search[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3128-3141.

[117] Zheng L, Wang S, Tian L, et al. Query-adaptive late fusion for image search and person re-identification[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1741-1750.

[118] Van de Weijer J, Schmid C. Applying color names to image description[C].IEEE International Conference onImage Processing, 2007. ICIP 2007. 2007, 3: III-493-III-496.

[119] Zhou W, Li H, Sun J, et al. Collaborative Index Embedding for Image Retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[120] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]. Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

[121] Jégou H, Chum O. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening[J]. Computer Vision–ECCV 2012, 2012: 774-787.

[122] Indyk P, Motwani R. Approximate nearest neighbors: towards removing the curse of dimensionality[C]. Proceedings of the thirtieth annual ACM symposium on Theory of computing. ACM, 1998: 604-613.

[123] 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.

[124] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]. 2006 IEEE computer society conference on Computer vision and pattern recognition . IEEE, 2006, 2: 2169-2178.

[125] 黄铁军, 施路平, 唐华锦,等. 多媒体技术研究:2015——类脑计算的研究进展与发展趋势[J]. 中国图象图形学报, 2016, 21(11):1411-1424.

[126] Chen Y, Luo T, Liu S, et al. Dadiannao: A machine-learning supercomputer[C].Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture. IEEE Computer Society, 2014: 609-622.

[127] Jouppi N P, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit[J]. arXiv preprint arXiv:1704.04760, 2017.

© 中国图象图形学报版权所有