一种用于图像检索的多层语义二值描述符

吴泽斌[1)] 于俊清[1),2)] 何云峰[1)] 管 涛[1)]

[1)](华中科技大学计算机科学与技术学院 武汉 430074)

[2)](华中科技大学网络与计算中心 武汉 430074)

摘 要 随着图像数据的爆炸性增长，基于内容的图像检索引起了大量的关注. 图像检索系统的性能很大程度上是由描述符决定的. 有很多传统的描述符先后被提出，但检索的准确率都不太理想. 随着深度学习的发展，利用卷积神经网络(Convolutional Neural Network, CNN)来学习占用空间小且具有较强区分力(discriminative)的图像表示逐渐兴起. 卷积神经网络全连接层的特征通常为分类任务而设计，捕获的往往是高层的语义信息，难以充分有效的捕获图像的局部信息，而且维度很高. 为解决全连接层特征缺乏局部信息且维度较高的问题，本文提出了一种多层语义二值描述符(Multi-level Semantic Binary Descriptor, MSBD). 多层语义二值描述符通过多层语义浮点描述符构建和二值描述符学习两个步骤生成. 多层语义浮点描述符由全局分支、对象分支以及显著性区域分支构成，每个分支代表一个语义层次，可以同时捕获全局特征以及显著的局部特征. 二值描述符学习算法通过一个迭代的过程减少二值化过程中的量化误差以及编码中的冗余信息，在压缩描述符的同时减少区分力的损失. 为了进一步提高查询的准确率，本文提出了一种不相似性度量函数. 此度量函数同时包含了哈希代表的视觉语义信息以及类级别的高层概念语义信息. 本文在该领域典型的数据集上对描述符进行了系统的对比实验，实验结果表明，多层语义二值描述符具有很强的区分力，查询准确率优于很多当前最先进的浮点描述符，在 Oxford5K 数据集上与目前最好的方法达到了相近的准确率，在 Paris6K 数据集上比已有的方法超过了约 4.3%，在 Holidays 数据集上比已有方法超过了约 2.1%.

关键词 图像表示；卷积神经网络；不相似性度量；图像检索；多层语义二值描述符

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2020.01641

Multi-level Semantic Binary Descriptor for Image Retrieval

WU Ze-Bin[1)] YU Jun-Qing[1),2)] HE Yun-Feng[1)] Guan Tao[1)]

[1)] (*Department of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan* 430074)

[2)] (*Center of Network and Computation, Huazhong University of Science and Technology, Wuhan* 430074)

Abstract As the explosive growing of the multimedia data on the Internet, finding an interesting image meeting the user query demand is becoming more and more difficult today, and content-based image retrieval, which aims to find the database images similar to a query image given by the user, is attracting increasing attention. The performance of an image retrieval system is largely decided by the image descriptor used. A lot of traditional shallow image descriptor building frameworks have been proposed, however, the accuracy they achieve on image retrieval benchmark datasets is not satisfying because of the limited representation ability of the shallow descriptors. With the advent of deep learning, making use of convolutional neural network to learn compact and discriminative representation has attracted considerable interest recently, because the learning ability of convolutional neural network is very strong given enough

收稿日期：2019-08-27；在线发布日期：2020-02-07. 本课题得到国家自然科学基金（61572211, 61173114, 61202300）资助. **吴泽斌**，博士研究生，主要研究领域为图像检索、深度学习, E-mail: zbwu@hust.edu.cn. **于俊清**（通信作者），博士，教授，主要研究领域为数字媒体处理与检索、多核处理器编程环境, E-mail: yjqing@hust.edu.cn. **何云峰**，博士，讲师，主要研究领域为数字媒体处理与检索. **管涛**，博士，副教授，主要研究领域为移动视觉搜索、增强现实、计算机视觉.

training data and supervision information. Many methods usually use the fully-connected layer feature to generate the representation for image retrieval, because the features from the fully-connected layers are relatively informative compared with the former layers. However, convolutional neural network is usually trained for classification task, and the features from fully-connected layers of convolutional neural network usually capture high-level semantic information and lack sufficient local characteristics of the input image, the discriminative ability of the image descriptor is affected by this reason. What's more, the features from fully-connected layers are usually not so compact and consume lots of storage, the scalability is limited. To address this problem, we propose a multi-level semantic binary descriptor building method which can capture global and salient local features simultaneously. Instead of a popular end-to-end approach, our binary descriptor building method is composed of two stages: multi-level semantic real-valued descriptor building and binary codes learning. The multi-level semantic real-valued descriptor is built from three streams: global stream, object stream and salient stream, each stream captures the information of one semantic level. The real-valued descriptor is usually high-dimensional and lots of redundancy exists in it, consuming a lot of storage resource. In the second stage of our method, an iterative learning algorithm is proposed to learn compact and discriminative binary codes by incorporating a sparsity constraint. The learning algorithm aims to minimize the hashing quantization loss and reduce redundancy in the codes, preserving the discriminative ability of the real-valued descriptor when compressing it so as to achieve both compact and discriminative codes for image retrieval. Moreover, a dissimilarity metric is proposed by simultaneously incorporating visual-level information in hash codes with class-level and high-level semantic information to further increase the query accuracy of retrieval. Extensive experiments on image retrieval benchmark datasets demonstrate that our descriptor is effective. We compare our method with both binary methods and real-valued ones, and prove that our binary descriptor is not only compact, but also discriminative, even outperforms many state-of-the-art real-valued representation on image retrieval task. As the experiments show, our method is on par with the state-of-the-art methods on Oxford5K dataset, and outperforms the state-of-the-art method by 4.3% on Paris6K dataset, by 2.1% on Holidays dataset, proving the effectiveness of our method.

Keywords image representation; convolutional neural network; dissimilarity metric; image retrieval; multi-level semantic binary descriptor

1 引 言

随着卷积神经网络的提出与发展，很多新的图像表示被提了出来. 文献[1]表明，将卷积神经网络(Convolutional Neural Network，CNN)在相关的数据集上微调后，全连接层特征在图像检索任务上的区分力很强. 然而，文献[1]也表明，全连接层特征的区分力不如最先进的基于局部特征[2]的浅层图像描述符. 全连接层通常是为图像分类任务而训练，捕获的往往是全局语义特征，缺乏局部信息. 为解决此问题，文献[3]提出用滑动窗口法来生成多尺度的分片(patch)，然后利用各个分片的特征来生成描述符，此描述符的区分力超过了很多最先进的浅层表示. 与全连接层不同的是，卷积层通常捕获的是局部的模式. 很多研究者利用卷积层特征来生成图像的表示[4-6]. 与全连接层特征相比，卷积层特征的维度通常不高，可以像浅层的局部特征一样用一些方法[7-9]将其聚合成一个图像描述符.

尽管浮点描述符在图像检索任务上取得了较高的准确率，但占用空间较大，含有不少冗余信息. 一些研究者便开始用二值化描述符或哈希编码[10-16]表示图像. 哈希编码占用空间小，可以节省很多存储空间，而且二值编码间的距离计算可以用比特操作快速地进行. 为了得到好的哈希编码，哈希学习算法应当能保持相似性，使相似的图像具有相似的哈希编码，尽可能保持原浮点特征间的局部邻域结构. 而且，如果哈希编码的每一位之间能够保持相互独立，哈希编码将包含更少的冗余信息，具有更

强的区分力. 然而，传统的哈希方法基本只能在一定程度上满足这些要求，因此，基于 CNN 的端到端(end-to-end)深度哈希方法正在引起越来越多的关注[17-25]. 利用深度神经网络来学习哈希码是一个具有挑战性的任务，因为二值哈希函数不可微，后向梯度传播算法不适用. 为解决此问题，研究者们通常将 *sign* 函数松弛为 *sigmoid* 函数或 *tanh* 函数. 在最近的工作中[19-20,24-26]，深度离散哈希被提了出来以解决此问题，并取得了优良的性能.

本文旨在解决以下问题：单一全连接层特征不能够有效地捕获图像的局部信息，而且通常维度较高，消耗较大的存储空间. 针对此问题，本文提出了一种占用空间小且具有强区分力的多层语义二值描述符. 多层语义二值描述符框架由两部分构成：浮点图像描述符构建和二值描述符学习. 二值描述符学习算法以浮点图像描述符作为输入. 为了让描述符捕获更多信息，浮点描述符由三个语义层次构成，可以同时捕获全局信息和显著的局部信息，使用对象检测方法和显著区域检测方法来捕获局部显著性特征. 物体检测方法被用来提取方形的区域，而显著性区域检测方法则被用来提取任意形状的显著性区域. 在浮点图像描述符的基础上，一个迭代的学习算法被用来学习信息丰富、区分力强的二值描述符. 多层语义二值描述符(Multi-level Semantic Binary Descriptor, MSBD)是一个二阶段方法，在图像检索任务上取得了较高的准确率.

本文的主要贡献如下：

（1）基于多层语义浮点描述符，提出了一个迭代的二值描述符学习算法，以学习一个占用空间小且信息含量丰富的二值图像表示，并在图像检索任务上验证了其有效性.

（2）提出了一个不相似性度量函数，同时融合了哈希编码间的不相似性以及高层的类级别不相似性，可以有效提高查询的准确率.

（3）MSBD 在多个常用的图像检索数据集上得到了较高的准确率，甚至超过了很多浮点型描述符，表明了其有效性.

本文其余部分的组织结构如下：第 2 节介绍了相关工作；第 3 节描述了 MSBD 的构建方法；第 4 节是实验与分析；第 5 节是结论部分.

2　相关工作

MSBD 主要与用于图像检索的图像表示有关，包括基于 CNN 的浮点型深度图像表示以及二值哈希表示.

基于 CNN 的浮点型图像表示　随着深度卷积神经网络[27]的产生，很多基于 CNN 的深度图像表示被提了出来. 早期的工作[1][3][28]简单地使用全连接层特征作为图像描述符，但是，这些方法的准确率都没有超过最先进的基于浅层特征的描述符，因为全连接层特征通常捕获的是全局信息，而缺乏局部信息. 文献[4-6]等利用卷积层特征来构建图像表示. 卷积层特征可以捕获很多局部特征，而且维度不高. 可以用各种方法将卷积层局部特征聚合为一个图像描述符，如和池化(sum-pooling)[4]，最大值池化(max-pooling)[6]以及 VLAD (Vector of Locally Aggregated Descriptors)聚合[29]. 以上的方法均为两阶段方法：首先在图像检索数据集上对 CNN 进行微调，然后提取某层的特征来构建图像描述符. 为了增加描述符的几何不变性，Reddy 等[30]利用物体检测方法来生成多种尺度的 patches, 用 CNN 提取 patch 的描述符，然后通过 max-pooling 这些 patch 描述符来得到全局描述符. Reddy 等还结合 ITQ(Iterative Quantization)[12]提取了一个二值化的版本. 本文的方法除了包含物体层的信息，还包含了全局层和显著性区域的信息. 文献[31-33]提出了端到端的图像表示学习框架. 文献[33]将一个 VLAD 层嵌入到网络中，此层的参数可以通过后向传播算法进行调整. 文献[31]则利用一个三分支网络来学习图像表示，并且利用一个图像区域生成网络来选择显著性的区域，可以有效地捕获丰富的局部信息. 与此不同的是，文献[32]利用了一个二分支网络来学习图像表示，并且使用 SfM(Structure from Motion)[34]来生成训练的图像对(pair)，使得在图像检索数据集上的无监督学习成为可能. 本文使用的浮点描述符构建方法由三个分支构成，通过串连“全局分支”、“对象分支”和“显著性分支”的表示来融合图片全局、方形物体区域以及显著性区域的信息以生成一个结构性的包含多个语义层的表示，本文使用的浮点描述符构建方法的三个语义分支虽然都使用了同一个特征提取网络，但不是一个“端到端”的方法，特征提取网络仅使用整幅图片进行训练. 当图片的背景比较混乱或含有一些其它的物体时，描述符的性能会受到影响. 为了处理这一问题，Kim 等[35]在 R-MAC (Region Maximum Activation of Convolutions)[6]框架中加入 Attention (注意力)机制，利用 Attention 机制计算各个区域的权值，再进一步融合全局信息，同时利用局部信息和全局信息以生成对上下文敏感的区域特征. Kim 等提出的方法可以同时学习 Attention 层和描述符，与此不同的是，

本文的方法没有使用 Attention 机制来对特征进行加权，而是对显著性区域信息生成了一个描述符，并通过串连将全局描述符和对象层描述符一齐融合在浮点描述符中，以得到一个显示包含多个语义层的结构化表示.

二值哈希表示 早期的传统哈希方法，使用随机投影作为哈希函数[10,36-38]. 局部敏感哈希[10]使得相似的图片对应的哈希拥有较高的概率发生碰撞. 然而，局部敏感哈希是一种不依赖于数据的哈希方法，准确率通常很低，需要较长的哈希码来保证准确率. 文献[11]提出了一种保持相似性的哈希，此种哈希方法利用点对间的距离形成的 *Laplacian* 矩阵来生成哈希函数. 通过保持点对间的相似性，可以在一定程度上保持原浮点特征间的邻域结构，生成区分力更强的哈希码. 文献[39]提出了一种超球体哈希，将空间上一致的点映射到同一个哈希码. 不同于基于随机投影的哈希和超球体哈希，文献[14]首先利用 *k-means* 量化器对特征进行量化，然后用量化得到的索引作为哈希码. 文献[13]提出了一种二值重构嵌入(Binary Reconstruction Embedding, BRE)方法来最小化欧氏距离与哈希汉明距离间的差异，而不是直接最小化二值量化误差. 与文献[13]类似，文献[40]提出了一种自适应二值量化方法来学习哈希函数. 文献[12]提出了一种迭代算法来交替更新编码及旋转矩阵，最小化量化误差. 为了处理学习长码的问题，文献[41]在投影矩阵上引入了稀疏约束，并且证明了其有效性. 引入稀疏约束减少了训练过程过拟合(over-fitting)的可能.

传统的哈希方法基本是基于手工(hand-crafted)特征的，特征与哈希函数并不是同时学习的，因此哈希函数并不是最优的. 随着深度卷积神经网络的提出，很多深度哈希(deep hashing)方法被提了出来. 深度哈希方法可以同时学习特征与哈希函数. 深度哈希方法可以被分为三类：无监督式(unsupervised)哈希，监督式(supervised)哈希以及半监督式(semi-supervised)哈希. 由于半监督方法与本文关系不大，在此处不阐述. 文献[21-22]提出了一种无监督式的深度哈希方法，并利用三种约束来学习更好的哈希码：（1）通过最小化量化误差来减小信息损失，提高编码的区分力;（2）使编码均匀分布，增大哈希码的信息含量;（3）将旋转不变性融合进学习过程. 文献[21]在训练的过程中不需要标签信息. 文献[17]交替学习哈希码和网络参数，并在学习过程中保持编码的独立性与均衡性，减少信息损失和信息冗余，以得到最优的哈希码. 为了更好地捕获局部区域信息，文献[42]提出了一种深度区域哈希(DRH, Deep Region Hashing)方法，此方法同时学习 ROI-pooling 层和区域哈希(region hashing)层.

DRH 的 Regions 是由 RPN(Region Proposal Net)或滑动窗口法来生成的. 每个 region 都会生成一个哈希码以用于检索. 查询时，DRH 同时使用了全局信息和局部信息：DRH 首先用全局 DRH(gDRH)作为初始查询，然后用局部 DRH(lDRH)对初始查询的结果再进行一次排序. Song 等将二值的 DRH 与浮点方法进行了比较，表明 DRH 甚至要优于浮点的 R-MAC 方法. DRH 类似于本文的方法，同时使用了全局信息和局部信息，只是本文的方法将全局信息和局部信息同时集成在一个全局描述符中，然后再用于学习哈希，且本文没有使用任何再排序(re-ranking)策略. 为了适应图片类标签不存在的情况，文献[43]利用 SfM[34]来生成用于训练的图像对，并取得了较高的准确率.

无监督式哈希在训练的时候不需要监督信息，然而，无监督哈希的准确率通常不够高，因此监督式哈希也得到了大量的研究. 文献[44]利用点对相似度矩阵来生成近似的哈希码，并同时用此哈希码及图片标签来监督网络的训练. 文献[18]提出利用一个三元组排序损失(triplet ranking loss)函数来指导网络学习特征与哈希码. 三元组排序损失函数能够在一定程度上保持输入的三元组间的排序关系. 三元组排序损失函数对各个三元组给予了相同的权重，这对多标签数据集并不是很适用，因为并不能很好地反映图片标签间的相似关系. 文献[45]提出了一个加权三元组排序损失函数来处理这一问题. 然而，图像检索数据集通常没有标签，而且在训练集较大时，三元组空间将非常大，训练会变得较为困难. 文献[46]提出了一种层次深度哈希(HDH, Hierarchical Deep Hash), 训练时输入的是二元组，而不是三元组，并同时利用卷积层和全连接层构建了一个两层的哈希函数. 层次深度哈希的损失函数由三个部分构成：单点(point-wise)分类损失，点对(pairwise)损失以及哈希损失. HDH 由两个语义层次的哈希构成，第二层利用第一层的特征来学习语义层次以及压缩率更高的哈希码. HDH 并没有将两个语义层融合在一个全局描述符中，而是在不相似性度量中融合了两个语义层的哈希，用两幅图片间的两层哈希码距离的加权平均来计算两幅图片间的不相似度，此外，文献[46]还提出了一个基于显著性程度的方法来计算权值. 本文使用的浮点描述符将三个语义层信息融合在一个全局描述符中，而 HDH

则将两个语义层的信息融合在不相似性度量中. 本文的哈希经过一个迭代学习算法后隐式地包含了多个语义层的信息. 本文的不相似性度量也是一个加权平均的形式，是哈希距离与类概率信息的融合，进一步丰富了语义层次.

与此类似，文献[47]提出了一种层次语义哈希，其哈希函数由语义级相似度以及哈希级相似度构成. 文献[48]则提出了一种语义保持哈希来同时学习哈希和分类任务，此方法仅使用了单点损失函数，但是很有效. 哈希函数一般使用 *sign* 函数来进行二值化，但 *sign* 函数并非连续函数，因此一般用 *sigmoid* 函数或 *tanh* 函数来近似，然而 *sigmoid* 函数和 *tanh* 函数会降低训练收敛的速度. 为解决此问题，文献[49]提出了一个伸缩(scaled) *tanh* 函数 $-tanh(\beta_t x)$ 来近似 *sign* 函数，当 $\beta_t \to \infty$ 时，此伸缩 *tanh* 函数能收敛到 *sign* 函数. 与使用近似函数不同的是，文献[50]等通过离散循环坐标下降法(discrete cyclic coordinate descent)来直接逐位学习哈希码，文献[50]在训练时同时使用了点对信息和类信息，利用哈希码来引导分类器的训练.

哈希方法使用 *sign* 函数来将浮点特征二值化，会造成量化损失. 为了处理此问题，基于量化的哈希学习方法被提了出来. Cao 等[51]将 PQ(Product Quantization，积量化)[52]融入到哈希学习中以处理量化损失问题，提出了 DQN (Deep Quantization Network，深度量化网络). Duan 等[53]利用基于 K-AutoEncoders (KAE, K 路自动编码器)的 MQ (Multi-Quantization，多量化器法)来代替 *sign* 函数，提出了 DBD-MQ(Deep Binary Descriptor with MutiQuantization，基于多量化器的深度二值描述符). Yu 等[54]将 PQ 作为一个 CNN 层嵌入到网络中，并提出了一个非对称 Triplet 损失函数，此方法称之为 PQN (Product Quantization Network，积量化网络). Klein 等[55]提出了一个 end-to-end 方法-DPQ (Deep Product Quantization,深度积量化). DPQ 是一个监督式的方法，码书和网络参数可以通过 BP 算法一起学习. DPQ 利用一个 central 损失来使同类的点相互靠近.

3　多层语义二值描述符(MSBD)

全连接层信息含量丰富，具有较强的区分力，但全连接层通常是为分类任务而设计，捕获的为高层语义信息，缺乏局部信息，而且维度通常较高. 为解决此问题，得到一个区分力强而空间占有量又不大的描述符，本文提出了一个多层语义二值描述符(Multi-level Semantic Binary Descriptor, MSBD). MSBD 由两个阶段构成：浮点描述符构建阶段和二值描述符学习阶段. 浮点描述符构建阶段通过包含多个语义层次的信息来提高描述符的区分力. 本文使用的浮点描述符是通过串连多个语义层描述符构建的一个结构化的全局描述符，三个语义分支的描述符都是利用同一张图片经过各种变换生成的，含有不少冗余信息，二值描述符学习阶段以此浮点描述符作为输入，进一步通过一个迭代算法来学习一个占用空间小又具有强区分力的二值图像表示，在压缩描述符的同时保持查询的准确率. 为了进一步提高查询的准确率，提出了一种不相似性度量函数，以同时包含哈希层的视觉信息和类级别的概念语义信息. 本文的哈希学习算法不是一个“端到端”的方法，特征提取与哈希函数不是一齐学习优化的. 这样得到的哈希码也许不是最优的，但本文的哈希学习算法参数很少，只有哈希码和旋转矩阵，比基于 CNN 的端到端方法要少得多，基于 CNN 的方法还要学习大量的网络参数，需要大量的训练数据. 本文的方法可以适用于训练集较少的情况.

3.1　浮点图像描述符构建

浮点描述符构建阶段旨在生成一个能够捕获多个语义层次的具有强区分力的浮点描述符. 浮点描述符的构建框架由三个分支构成，可以分别捕获全局层(global-level)，对象层(object-level)以及显著性区域层(salient region-level)等三个语义层次的特征.

全局分支　全局分支用于捕获整幅图片的全局级别信息. 为了增强描述符的尺度不变性，此分支使用了一个多尺度策略. 图片通常含有一些细微的结构，在较大的尺度可以更好地捕获这些细节结构信息. 令 $\boldsymbol{I}$ 表示输入图片，则全局分支的表示的生成过程如下：

$$\begin{aligned} \boldsymbol{I}_i &= resize_i(\boldsymbol{I}) \\ \boldsymbol{g}_i &= FEN(\boldsymbol{I}_i) \\ \boldsymbol{F}_g &= (\sum_i^3 \boldsymbol{g}_i)/3 \end{aligned} \tag{1}$$

*FEN(Feature Extracting Net)*表示特征提取网络. $\boldsymbol{I}_i$ 表示经过缩放后的第 *i* 个尺度的图片，$\boldsymbol{g}_i$ 表示第 *i* 个尺度图片的描述符，$\boldsymbol{F}_g$ 表示全局分支的描述符. 特征提取网络用于提取图片或 patch 的特征，在三个分支中都会用到，且三个分支中的特征提取网络是同一个，其架构如图 1 所示. 此网络将在后面的小节中详细介绍. 首先，将输入图片缩放到三个尺度{1.25, 1, 0.75}，然后将此三张图片输入到特征提取网络（*FEN*）中，最后对生成的三个特征($\boldsymbol{g}_i$)求平均，并进行 2-范数标准化(*L2-normalization*)，便

得到了包含多尺度信息的全局分支描述符 $\boldsymbol{F}_g$.

对象分支 由于 CNN 全连接层并不能很好地捕获局部物体信息，因此引入对象层来捕获这些信息. 首先，通过对象检测器来检测输入图片中的物体. 然后，选择信息含量最为丰富的若干图像分片(patch)，将它们缩放到 340×340，并输入到特征提取网络中，提取各个分片的特征. 最后，将这些分片特征聚合成一个描述符. 此处并没有使用常用的max-pooling和sum-pooling，而是像 MOP_CNN (Multi-scale Orderless Pooling)[3]方法一样使用了 VLAD 方法，以便更好地捕获各分片的局部信息. 在生成 VLAD 时使用了内部标准化(intra-normalization)[56]和软分配(soft-assignment)[57]，每个特征被分配到 10 个码字，码书的大小设为 200. 令 $\boldsymbol{I}$ 表示输入图像，则对象分支的表示的生成过程如下：

$$\begin{aligned}&\{\boldsymbol{p}_i\}=D_t(resize(\boldsymbol{I}))\\&\boldsymbol{f}_i=FEN(\boldsymbol{p}_i)\\&\boldsymbol{F}_o(\boldsymbol{I})=VLAD\_Pooling(\{\boldsymbol{f}_i\})\end{aligned}\tag{2}$$

D_t 为对象检测器，用于检测对象并选择图像分片. FEN 是特征提取网络，$\boldsymbol{p}_i$ 表示第 i 个选择的图像分片，$\boldsymbol{f}_i$ 表示第 i 个选择的图像分片的描述符，$\boldsymbol{F}_o(\boldsymbol{I})$ 是对象分支的表示. 此处的方法类似于同样使用了多尺度分片的 MOP_CNN，但是此处的方法基于 MOP_CNN 进行了几点改进：（1）MOP_CNN 用的是滑动窗口法，而非对象检测器. 滑动窗口法生成的分片中有很多来自于图片背景，并不含有多少有用的信息，对描述符的区分力甚至会有损害. （2）MOP_CNN 的各个分片的特征来自于 CNN 的全连接层，缺乏局部信息，而此处的各个分片的特征信息提取自特征提取网络的多个卷积层，可以捕获更多的局部信息. （3）MOP_CNN 为每个尺度生成一个描述符，并将它们串连起来，而本文使用的浮点描述符在全局分支用的是平均法，在物体分支则仅为所有的分片生成了一个 VLAD. 为了充分捕获各个尺度的对象信息，此处先将输入图片缩放到 3 个尺度{1.25, 1, 0.75}，然后再将它们送入对象检测器. 对象分支的表示的生成过程如下：

$$\begin{aligned}&\{\boldsymbol{p}_i\}=\bigcup_s(D_t(resize_s(\boldsymbol{I})))\\&\boldsymbol{f}_i=FEN(\boldsymbol{p}_i)\\&\boldsymbol{F}_o(\boldsymbol{I})=VLAD\_Pooling(\{\boldsymbol{f}_i\})\end{aligned}\tag{3}$$

对每一张输入图片，选择的分片数是 50. 在缩放图片时，保持图片的长宽比不变. 随着对象检测的发展，有很多对象检测器被提了出来，如 Selective Search[58], EdgeBoxes[59], MCG(Multi-scale Combinatorial Grouping)[60]. 随着深度学习的发展，很多基于 CNN 的对象检测器[61, 62]被提了出来. 然而，基于 CNN 的对象检测器除了需要类级别的监督信息外，通常还需要物体边界方框(bounding box)监督信息. 因此，此处选用的是 Edgebox.

显著性区域分支 显著性区域分支用于捕获显著性区域的特征. 对象检测器生成的分片是方形的，这些分片有可能只包含了对象的一部分，即一个对象有可能被截断了. 显著性区域检测器可以有效捕获图片的前景(foreground)部分，不是方形的，而且可以去除有干扰性的背景信息. 此处首先将输入图片缩放到 500×500，然后输入到显著区域检测器中，生成图片的显著性图(saliency map). 显著区域分支的表示 $\boldsymbol{F}_s$ 的生成过程如下：

$$\begin{aligned}&\boldsymbol{M}=resize(S(\boldsymbol{I}))\\&\boldsymbol{R}=\boldsymbol{M}\otimes\boldsymbol{I}\\&\boldsymbol{F}_s=FEN(\boldsymbol{R})\end{aligned}\tag{4}$$

S 表示显著性区域检测器，$\boldsymbol{M}$ 是生成的显著性图，$\boldsymbol{R}$ 是图片的显著性区域. 将显著性图缩放到 500×500，然后与输入图片进行点乘，便得到了显著性区域 $\boldsymbol{R}$. 最后，将显著性区域输入到特征提取网络中，便得到了显著性区域分支的表示. 此处使用文献[63]中的方法作为显著性区域检测器，此检测网络基于 VGGNet[64]，从 HED(Holistically-nested Edge Detection)边缘检测器[65]发展而来，利用了多层和多尺度信息，获得了优良的显著性对象检测性能. 由于图像检索数据集没有提供边界方框信息，因此此处直接使用了文献[63]中预训练(pre-trained)的网络，而且通过实验表明了此分支信息的有效性.

在得到三个分支的表示后，将它们融合到一个描述符中，记为 $\boldsymbol{F}_{MSBD}$. 具体过程如下：

$$\begin{aligned}&\boldsymbol{F}_1=Normalize(\boldsymbol{P}_1\cdot\tilde{\boldsymbol{F}}_{\mathrm{g}})\\&\boldsymbol{F}_2=Normalize(\boldsymbol{P}_2\cdot\tilde{\boldsymbol{F}}_{\mathrm{o}})\\&\boldsymbol{F}_3=Normalize(\boldsymbol{P}_3\cdot\tilde{\boldsymbol{F}}_{\mathrm{s}})\\&\boldsymbol{F}_{MSBD}=[\boldsymbol{F}_1,\boldsymbol{F}_2,\boldsymbol{F}_3]\end{aligned}\tag{5}$$

首先将各个分支的表示进行 2-范数标准化，然后进行 PCA，并再次进行标准化(Normalize). $\tilde{\boldsymbol{F}}_{\mathrm{g}}$ 是标准化后的全局分支的表示，$\tilde{\boldsymbol{F}}_{\mathrm{o}}$ 是标准化后的对象分支的表示，$\tilde{\boldsymbol{F}}_{\mathrm{s}}$ 是标准化后的显著性区域分支的表示. $\boldsymbol{P}_1$，$\boldsymbol{P}_2$，$\boldsymbol{P}_3$ 分别是三个分支描述符的 PCA 投影矩阵. 全局分支的每一个尺度的描述符首先进行 2-范数标准化，利用 PCA 降维到 $D1$，再次标准化，最后求平均. 此处通过串连三个分支的描述符来融合三个层次的语义. 为了进一步增强整个图像表示的区分力，还串连上了 Hessian-Affine-rootSIFT-

VLAD[66]，码书大小设为 64. 在后面的各节中，将 $\boldsymbol{F}_{\boldsymbol{MSBD}}$ 称为“MSBD-float”.

特征提取网络　特征提取网络用于提取图片、分片的深度特征. 特征提取网络的架构如图 1 所示. 去掉了此网络的全连接层，以便使其适用于任意大小的输入图片. 使用了特征提取网络的多个卷积层来提取特征，因为不同的层可以捕获不同抽象级别的模式区域. 将各个层的卷积层特征进行池化、串连、2-范数标准化，便得到了用此网络要生成的特征. 类似于文献[6]，此处使用了全局最大值池化(global max-pooling，GMP)分别处理各个卷积层的特征. 特征提取网络要先在各个要测试的数据集上进行微调，并在后面增加一个全连接层和一个 *softmax* 层作为一个分类器，全连接层的大小就是各个数据集的类数. 在训练特征提取网络时，将输入图片缩放到 513×513，以 VGGNet-16 作为特征提取网络的骨架网络，使用 conv5_1, conv5_3, pool5 来提取特征，仅微调 conv4_3 以后的各层，在全局最大值池化(global max-pooling)层前面增加了一个 L2-normalization 层. 各个数据集的训练参数设置均不一样.

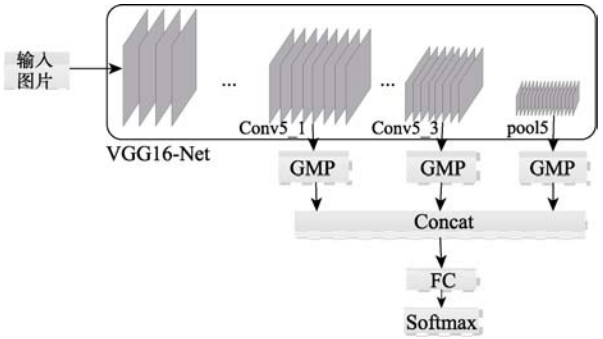


图 1　特征提取网络架构

3.2　二值描述符学习

二值描述符学习过程以浮点描述符构建过程生成的浮点描述符作为输入，生成一个占用空间小、且具有强区分力的二值图像表示，在压缩图像表示的同时保持其查询的准确率. 为了使生成的二值表示信息含量丰富、区分力强，通常需要在学习的过程中施加一些约束条件，如：(1) 独立性. 为了保持更多的信息，减少信息冗余，应当使各位之间尽可能相互独立. (2) 均衡性. 使各位具有相等的概率为 0 或 1. 除此之外，浮点表示与二值表示之间的量化损失应当尽可能的小. 为了满足这些条件，本文提出了一个二值描述符学习算法，在减少量化损失的同时减少信息冗余. 该学习算法的目标函数如下：

$$\begin{aligned} &f(\boldsymbol{X};\boldsymbol{B},\boldsymbol{R})=\min_{R,B}\left\|\boldsymbol{RX}-\boldsymbol{B}\right\|_2+\partial\left\|\boldsymbol{B}\right\|_1 \\ &\boldsymbol{RR}^{\mathbf{T}}=\mathbf{I} \\ &\left|\boldsymbol{b}_i\right|\leqslant k \end{aligned} \tag{6}$$

此目标函数的第一项表示的是量化损失. $\boldsymbol{X}$ 是输入，$\boldsymbol{R}$ 是正交旋转矩阵，用于旋转输入向量 $\boldsymbol{X}$，使其各维间的独立性更强. 另一方面，$\boldsymbol{R}$ 也用于降低 $\boldsymbol{X}$ 的维度，使其维度与二值表示的位数相等. 第二项是一个稀疏约束，k 用于控制二值表示中 1 的个数，即二值表示的稠密程度. k 越小，二值表示越稀疏. α 被设为 0.001. 使用坐标下降法(Coordinate Descent)来交替优化 $\boldsymbol{B}$ 和 $\boldsymbol{R}$. 第一步，初始化 $\boldsymbol{B}$ 和 $\boldsymbol{R}$. 用 PCA 旋转矩阵来初始化 $\boldsymbol{R}$，而不是用一个随机矩阵，可以在降维的同时尽可能保留更多的信息. 此处并没有用 $sign(\boldsymbol{RX})$函数来初始化 $\boldsymbol{B}$，而是使用 ITQ 方法[12]来初始化 **B**，以便使编码有更好的初值，ITQ 的迭代次数设为 50. 第二步，根据 k 值来稀疏化 $\boldsymbol{B}$. 采取逐列稀疏化 $\boldsymbol{B}$ 的方式，$\boldsymbol{B}$ 的每一列对应一个输入向量. 对于每一列 $\boldsymbol{b}_i$，计算其包含的 1 的个数，如果 1 的数目大于 k，则将一些 1 置为 0，在置 0 时要保证量化损失不增加. 第三步，固定 $\boldsymbol{B}$，优化 $\boldsymbol{R}$. 这是一个正交 Procrustes 问题[12]. 令 $\boldsymbol{C}=\boldsymbol{BR}^{\boldsymbol{T}}$，然后使用奇异值分解：$\boldsymbol{USV}^{\boldsymbol{T}}=svd(\boldsymbol{C})$，$\boldsymbol{R}=\boldsymbol{UV}^{\boldsymbol{T}}$. 第四步，固定 $\boldsymbol{R}$，优化 $\boldsymbol{B}$. 最后，迭代后面这三步. 详细过程如**算法 1** 所示. 此算法可以看作是 ITQ 的一个改进算法，在 ITQ 的基础上加入了稀疏性约束，在迭代算法中加入了稀疏化操作，以减少冗余信息. 此外，本文没有使用汉明距离，而是类似于 HDH[46]，提出了一个不相似性度量函数，通过加权同时融合了哈希不相似性度和类概率信息.

3.3　算法复杂度分析

算法 1 的复杂度为 $O(m_0\times L\times maxIter\times N)$，影响此算法复杂度的因子主要有以下几个方面：(1) 训练集大小 N；(2) 最大迭代次数 $maxIter$；(3) 码长 L. m_0 是置 0 方案总数，设为常数. 此算法的复杂度与训练集的大小成正比，其它几个因子都远小于 N.

3.4　不相似性度量函数

两个二值编码的不相似性通常用汉明距离来表示.

算法 1　二值描述符学习算法

输入：浮点描述符向量矩阵 $\boldsymbol{X}$，与稠密度相关的因子 k，码长 L，最大迭代次数 $maxIter$

输出：旋转矩阵 $\boldsymbol{R}$，二值描述符 $\boldsymbol{B}$

1. 用 PCA 矩阵初始化 $\boldsymbol{R}$，旋转 $\boldsymbol{X}$

2. 用 ITQ 方法初始化 $\boldsymbol{B}$;

3. FOR iteration t=1 to $maxIter$

// 使其稀疏化;

4. FOR i=1 to N

5. $\boldsymbol{b_i}$是 $\boldsymbol{B}$ 的第 i 列;

6. 计算 $\boldsymbol{b_i}$中 1 的个数: n_1

7. IF $\{n_1 < k\}$

8. 从 $\boldsymbol{b_i}$中随机选择 k-n_1 个值为 1 的位置; 将这个过程重复 1000 次，得到 1000 种置 0 的方案;

9. 计算每种置 0 方案的量化损失，选择使量化损失最小的方案，将此最小的量化损失记为 loss_q;

10. IF{ loss_q 比置 0 前的量化损失小}

11. 采用此量化策略对 $\boldsymbol{b_i}$进行稀疏化

12. END IF

13. END IF //n_1<k

14. END FOR // i=1 to N

15. 固定 $\boldsymbol{B}$，优化 $\boldsymbol{R}$: 使用正交 Procrustes 问题的解法;

16 旋转输入特征矩阵 $\boldsymbol{X}$: $\boldsymbol{X}$=$\boldsymbol{RX}$

17 固定 $\boldsymbol{R}$,优化 $\boldsymbol{B}$: $\boldsymbol{B}$=$\boldsymbol{sgn}$($\boldsymbol{RX}$);

18 END FOR //iteration

度量，然而汉明距离空间是一个整数空间，其空间大小为 L+1, L 是码长，即不同的汉明距离数很少. 为了增加距离度量的区分力，本文的不相似性度量同时融合了汉明距离以及一个类级别的不相似性，以便能同时利用类级别的语义. 此处的类级别不相似性是用 CNN 的 *softmax* 分类器输出的概率向量来计算的，类概率向量表示的是输入图片与各个类的相关程度，或者说属于各个类的程度. 概率向量的大小是由数据集的类数决定的，一般不大. 此处使用 ResNet-50[67]来得到所有图片的类概率向量. 提取类概率向量的过程可以离线进行. ResNet-50 首先要在各个数据集上微调.

令 $\boldsymbol{I_1}$, $\boldsymbol{I_2}$ 表示两张图片，$\boldsymbol{p_1}$, $\boldsymbol{p_2}$ 表示其对应的 *softmax* 输出的概率向量，$\boldsymbol{H_1}$, $\boldsymbol{H_2}$ 是相应的二值图片表示. 以 $\boldsymbol{I_1}$为查询图片,则 $\boldsymbol{I_1}$, $\boldsymbol{I_2}$间的不相似性 $D(\boldsymbol{I_1}, \boldsymbol{I_2})$可以如下计算:

$$D(\boldsymbol{I_1}, \boldsymbol{I_2}) = (1 - W_{i,j}) \cdot D_p(\boldsymbol{p_1}, \boldsymbol{p_2}) + D_h(\boldsymbol{H_1}, \boldsymbol{H_2}) \quad (7)$$

$D_p(\boldsymbol{p_1}, \boldsymbol{p_2})$表示的是用类概率向量计算的不相似性，本文用的是 L1 距离; $D_h(\boldsymbol{H_1}, \boldsymbol{H_2})$是二值表示间的汉明距离. $W_{i,j}$是一个权重，表示的是 $\boldsymbol{I_1}$和 $\boldsymbol{I_2}$属于同一个类的概率. $W_{i,j}$的计算公式如下:

$$W_{i,j} = < \boldsymbol{p_1}, \boldsymbol{p_2} > \cdot max(\boldsymbol{p_2}) \quad (8)$$

$<\boldsymbol{p_1}, \boldsymbol{p_2}>$表示 $\boldsymbol{p_1}$, $\boldsymbol{p_2}$间的内积,$max(\boldsymbol{p_2})$是 $\boldsymbol{I_2}$的类概率，以概率向量各维中最大的值作为类概率.

4 实 验

4.1 数据集

INRIA Holidays 数据集[68]有 1491 张图片. 这个数据集有 500 组图片，每组图片包含同一个物体，或者属于同一个场景. 每组图片都有一张作为查询图片，剩下的 1490 张则作为数据库图片. 检索性能用 mAP (mean average precision, 平均准确率)度量. mAP 是准确率-召回率曲线与坐标轴围成的面积.

Oxford5K 建筑物数据集[69]有 5062 张 Oxford 建筑物的图片. 11 类地标建筑物的 55 张图片被用作查询图片. 该数据集提供了查询图片的物体边界方框(bounding box)，但本文没有使用这个边界方框，而是将整张查询图片作为网络的输入. 检索性能用 mAP 度量. 本文还在 Oxford105K 数据集(Oxford5K+100K 干扰图片)上做了实验.

Paris6K 数据集[69]有 6412 张 Paris 地标图片. 类似于 Oxford5K 数据集，11 类地标的 55 张图片被用作查询图片，此数据集也提供了查询图片的物体边界方框(bounding box)，但本文也没有用. 此数据集的性能用 mAP 度量.

Baselines 本文将 MSBD 二值描述符以及其变体 MSBD-float, MSBD-h, MSBD-s 与非监督式方法、监督式方法以及浮点描述符进行了对比. MSBD-h 和 MSBD-s 是用于进行消融研究(Ablation study)的，以用于验证不相似性度量及稀疏约束的有效性. MSBD-float 是二值化之前的浮点描述符，是二值化算法的输入. MSBD-float 使用 2-范数距离作为不相似性度量. MSBD-h 使用汉明距离作为不相似性度量，而没有使用本文提出的不相似性度量. MSBD-s 在二值化学习时没有使用稀疏性约束，但使用了本文的不相似性度量. MSBD 与非监督式的哈希方法进行了比较，包括传统的哈希方法：ITQ[12], KMH (K-Means Hashing)[14], SH(Spectral Hashing)[11]，以及基于 CNN 的深度哈希方法：DeepBit[22], DRH[42], P2B (Pixels to Binary codes)[43]. MSBD 也与监督式哈希方法进行了比较：SSDH(Supervised Semantics-preserving Deep Hashing)[48]和 HDH[46]. 而且，还与浮点描述符进行了比较：Neural codes[1], SpoC (Sum-Pooled Convolutional features)[4], R-MAC[6], CroW (Cross-dimensional Weighting)[5], Faster-RCNN[70], NetVLAD[33]以及 DIR (Deep Image Retrieval)[31].

4.2　实验细节

使用 Caffe 训练 CNN. 对所有的图片进行了缩放，使大边为 1024，保持长宽比不变. Hessian-Affine-rootSIFT-VLAD 以及每一个分支的维度，即 $D1$，被设为 512. 在 Oxford5K 数据集和 Paris6K 数据集上，特征提取网络输入批量的大小(batch size)设为 16，初始学习率设为 0.01，SGD 最大迭代次数设为 10000. 在 Holidays 数据集上，初始学习率设为 0.001，并在 L2-normalization 层前增加了一个 *BatchNorm* 层[71]，最大迭代次数设为 5000.

4.3　二值描述符学习算法的收敛性

二值描述符学习算法的收敛性如图 2 所示. 在图 2 中描述了误差函数随迭代次数的变化，以观察二值描述符学习算法的收敛性. 二值描述符的稠密度(k/L)设为 0.3. 误差函数轴使用的是对数尺度. 从图 2 可以看出来，二值描述符学习算法具有良好的收敛性，在两个数据集上，误差函数在前 10 次迭代时下降都很迅速，然后逐渐趋于缓和. 从图 2 可以看出，二值描述符学习算法在 50 次时已经收敛，因此，在后面所有的实验中将最大迭代次数设为 50.

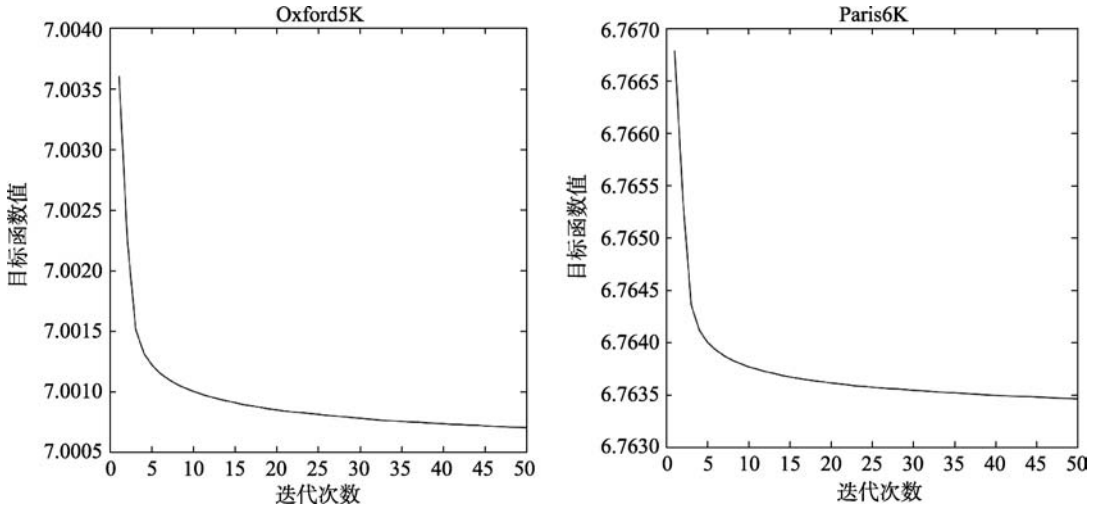


图 2　二值描述符学习算法的收敛性（此图表明的是误差随迭代次数的变化. 左边的是 Oxford5K 数据集，右边的是 Paris6K 数据集）.

4.4　算法对参数的敏感性

k 表示的是二值描述符中 1 的个数，二值描述符的稠密度为 k/L，L 是码长. 稠密度越大，二值描述符中 1 的个数越多. 在此处，将码长设为 256.

准确率对二值描述符稠密度的敏感性如图 3 所示. 从图 3 可以看出，mAP 在 Oxford5K 数据集和 Holidays 数据集上随 k 值有一定程度的波动，但在 Paris6K 数据集上很稳定. Holidays 数据集上的 mAP 在稠密度为 0.2 时达到最大，然后下降. 在后面的实验中，如果不特别说明的话，在 Oxford5K 和 Paris6K 数据集上，将稠密度设为 0.3，在 Holidays 数据集上设为 0.2.

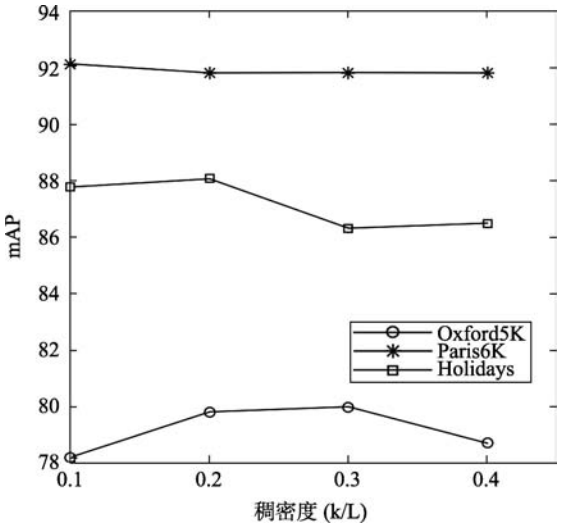


图 3　准确率对二值描述符稠密度的敏感性

4.5　MSBD与传统哈希方法的比较

MSBD 与无监督的哈希方法的比较如图 4 所示. MSBD 与无监督的哈希方法进行了比较. 无监督的哈希学习算法以 MSBD-float(MSBD 对应的浮点描述符)作为输入，使用汉明距离计算不相似度. 从图 4 可以看出，MSBD 在短码情况下要优于 ITQ, KMH 和 SH. 在 Oxford5K 数据集上，所有方法的 mAP 都随码长的增加而增加，ITQ 与 MSBD 的准确率差距在 32 位时最大，而且 MSBD 的准确率在所有的码长情况下都超过了其它的方法，甚至比 KMH

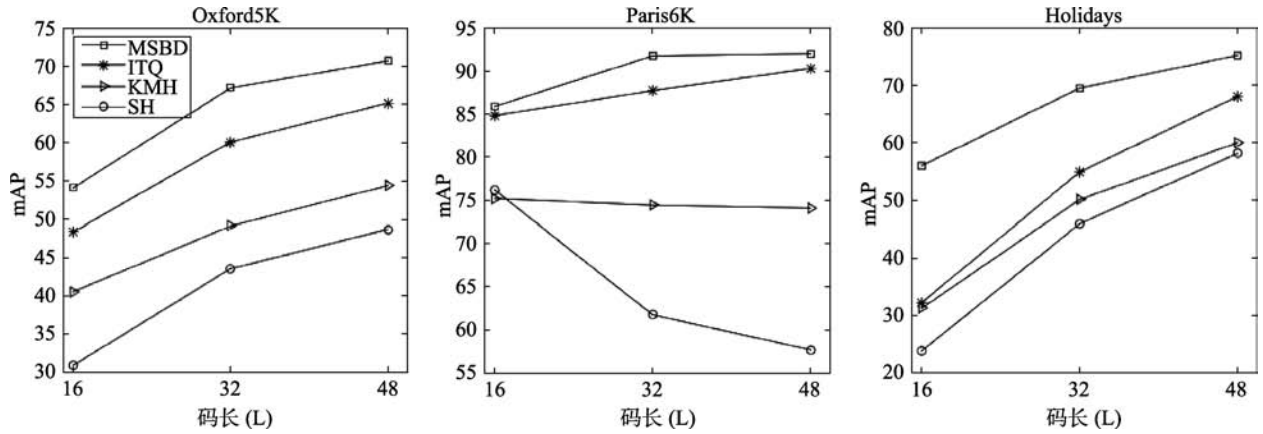


图 4 MSBD 与传统的哈希方法在不同码长时的准确率的对比. 从左到右依次是 Oxford5K, Paris6K, Holidays 数据集

和 SH 要高出 10 个百分点. 在 Paris6K 数据集上，除了 ITQ 外，其它方法的准确率并不是随着码长的变化而单调变化的，MSBD 的 mAP 在 32 位时达到最大，之后随着码长的增加减小了一点，因为噪声随着码长的增加也增加了. 在 Holidays 数据集上，KMH 的 *nsubit* 参数在 48 位时设为 3. Holidays 数据集上的情况与 Oxford5K 类似，mAP 随着码长的增加而增加，MSBD 的 mAP 远远优于其它方法，其它方法与 MSBD 之间的差距在 16 位时达到最大.

4.6 MSBD与端到端的深度哈希方法的比较

MSBD 与其变体以及端到端(end-to-end)的深度哈希(deep hashing)进行了对比，包括无监督式哈希和监督式哈希. MSBD 先与无监督式哈希进行对比，然后与监督式哈希进行了对比.

由于在 Oxford5K 及 Paris6K 数据集上学习深度哈希的方法并不多，在与无监督式哈希方法进行对比时，将 MSBD 与 DeepBit[22]以及 DRH[42]进行了比较，这二种方法也在图像检索数据集上做了实验，而且是当前最先进的无监督式哈希方法之一。此外，还与 ODFP(Object-level Deep Feature Pooling)+ITQ[30][22]进行了对比，二值化 ODFP 的 mAP 引用自文献[22]。ODFP 同样使用了物体检测方法来捕获物体层的信息，ODFP+ITQ 则表示用 ITQ 对 ODFP 进行二值化。MSBD 与无监督哈希在 Paris6K、Oxford5K 上的比较如表 1、表 2 所示. 从表 1 和表 2 可以看出，ODFP+ITQ 远不如 MSBD. DeepBit 方法不仅使用了量化损失，还使用了二值化损失以使编码趋近于二值化. 此外，DeepBit 还在编码上加了一个几何不变性约束，使经过几何变换后的图片的哈希码与原图的哈希码尽可能相似. DeepBit 没有使用图片类标签信息. 但是，从表 1 可以看出，MSBD 在各个码长下的准确率都要高于 DeepBit，在码长为 256 位和 512 位时，MSBD 在 Paris6K 数据集上的准确率比 DeepBit 分别要高 9.28%和 7.55%. DRH 方法使用了一个区域生成网络(Region Proposal Net)来生成感兴趣区域，并使用了区域哈希. 为了进行对比，使用的是没有用查询扩展的 gDRH 版本. 在 Paris 数据集上，MSBD 在 256 位与 512 位时的准确率分别比 DRH 高出 28.88%和 18.05%. 实际上，MSBD 在 512 位时的 mAP 比同时使用了 1024 位的全局 DRH(gDRH) 和局部 DRH(lDRH) 的方法 (gDRH+lDRH)[42]的 mAP(Paris: 0.801，Oxford: 0.783)还要高(分别高 10.35%，1.4%)，且局部 DRH(lDRH)是一个集合(每个 region 一个)，lDRH+gDRH 的总位数要远超 MSBD. MSBD 还与 DeepBit, DRH 以及最近提出的 P2B 方法[43]在 Oxford5K 数据集上进行了比较. 从表 2 可以看出，MSBD 的 mAP 要远远超过 DeepBit 和 DRH，甚至超过了 10 个百分点. 此外，在 256 位和 512 位时，MSBD 在 Oxford5K 数据集上的准确率甚至比 P2B 高出了 10.75%和 4.90%，进一步证明了本方法的有效性. P2B 使用 SfM 来生成训练时的匹配对(matching pairs)与不匹配对，然而，这样生成的匹配对集合中含有大量的噪声，对检索的性能产生了负面影响. 从表 1 和表 2 可以看出，MSBD-h 在没有使用汉明距离的情况下，依然要优于其它方法，证明了本文哈希学习算法的有效性；而 MSBD

表 1 MSBD 与无监督式哈希在 Paris6K 数据集上的比较

	256	512
ODFP+ITQ[30][22]	67.1	73.9
DeepBit[22]	82.50	82.90
DRH[42]	62.90	72.40
MSBD-h	89.72	88.94
MSBD-s	91.23	89.75
MSBD	**91.78**	**90.45**

表 2　MSBD 与无监督式哈希在 Oxford5K 数据集上的比较

	256	512
ODFP+ITQ[30][22]	48.9	50.8
DeepBit[22]	60.30	62.70
DRH[42]	58.30	66.80
P2B[43]	69.20	74.84
MSBD-h	77.39	77.44
MSBD-s	78.81	78.47
MSBD	**79.95**	**79.74**

要优于 MSBD-h 和 MSBD-s，证明了本文的不相似性度量以及稀疏约束的有效性.

在与监督式哈希方法进行比较时，MSBD 与 SSDH[48]以及 HDH[46]进行了比较. SSDH 以及 HDH 也在图像检索数据集上做了实验，并且是当前最先进的监督式方法之一. MSBD 与监督式哈希在 Paris6K、Oxford5K 上的比较如表 3、表 4 所示. 从表 3 可以看出，在 Paris6K 数据集上，MSBD 的准确率在 512 位时比 SSDH 要高出 6.58%，在 256 位与 512 位时比 HDH 分别要高出 6.58%和 3.15%. SSDH 同时学习高层的类语义以及哈希码，很大程度上捕获的是全局语义信息，缺乏中层以及低层的语义信息. HDH 与 MSBD 类似，也使用了多个语义层的信息. HDH 包含两个哈希层，分别代表不同的语义层次，HDH 的不相似度是这两层哈希的不相似度的加权平均. MSBD 不仅本身包含了多个语义层次的信息，而且，还在不相似性度量中融合了哈希代表的视觉信息和类概率向量代表的高层语义信息. 从表 4 可以看出，在 Oxford5K 数据集上，SSDH，HDH 与 MSBD 的准确率差距甚至比在 Paris 数据集上的更大，在 256 位和 512 位时，MSBD 的准确率比 HDH 分别要高出 10.25%, 9.24%，差距达到了 10 个百分点.

表 3　MSBD 与监督式哈希在 Paris6K 数据集上的比较

	256	512
SSDH[48]	–	83.87
HDH[46]	85.20	87.30
MSBD-h	89.72	88.94
MSBD-s	91.23	89.75
MSBD	**91.78**	**90.45**

表 4　MSBD 与监督式哈希在 Oxford5K 数据集上的比较

	256	512
SSDH[48]	–	63.80
HDH[46]	69.70	70.50
MSBD-h	77.39	77.44
MSBD-s	78.81	78.47
MSBD	**79.95**	**79.74**

4.7　MSBD与浮点图像描述符的比较

MSBD 与浮点描述符在 Oxford5K, Paris6K 以及 Holidays 数据集上进行了比较，还在 Oxford5K 数据集做了消融研究(ablation study)，以证明融合各个语义层次的有效性，在做消融实验时，将 $D1$ 设为 256. MSBD-float$_g$ 是全局分支的浮点描述符，MSBD-float$_o$ 是对象分支的浮点描述符，MSBD-float$_s$ 是显著性区域分支的浮点描述符. MSBD-float$_{go}$ 融合了全局(global)分支与对象(object)分支，MSBD-float$_{gs}$ 融合了全局(global)分支与显著性(salient)分支，MSBD-float$_{os}$ 融合了对象(object)分支与显著性(salient)分支. MSBD-float 是融合了三个分支的描述符. 在 Oxford5K 和 Paris6K 数据集上，“MSBD (binary)”的 $D1$ 被设为 512. MSBD 与浮点描述符在 Oxford5K 和 Oxford105K 数据集上的比较如表 5 所示. 从表 5 可以看出，MSBD-float 的准确率要高于其变

表 5　MSBD 与浮点描述符在 Oxford5K 和 Oxford105K 数据集上的比较

	D	Oxfd5K	Oxfd105K
TE+DA[73]	8064	67.6	61.1
FAemb[74]	15525	70.9	–
Nerual codes[1](float)	256	55.7	–
SpoC[4] (float)	256	58.9	50.1
R-MAC[6] (float)	512	66.9	61.6
CroW[5] (float)	512	70.8	63.2
Faster-RCNN[70](float)	4096	71.0	–
NetVLAD[33](float)	256	63.5	–
RADF[35](float)	2048	76.8	73.6
CIR[32](float)	512	80.1	75.1
DIR[31](float)	512	**83.1**	**78.6**
MSBD-float$_g$	256	66.7	–
MSBD-float$_o$	256	66.1	–
MSBD-float$_s$	256	60.6	–
MSBD-float$_{go}$	512	72.3	–
MSBD-float$_{gs}$	512	0.719	–
MSBD-float$_{os}$	512	0.743	–
MSBD-float(D1=256)	1024	76.5	–
MSBD-float(D1=512)	2048	78.6	70.2
MSBD(binary)	128	77.6	–
MSBD(binary)	256	80.0	–
MSBD(binary)	1024	–	67.52
MSBD(binary)	2048	–	69.62

体 MSBD-float$_g$，MSBD-float$_o$，MSBD-float$_s$ 以及 MSBD-float$_{go}$，MSBD-float$_{gs}$，MSBD-float$_{os}$ 证明融合多个层次语义信息的有效性. 从表 5 可以看出，在 Oxford5K 数据集上，MSBD 的准确率超过了大部分方法，包括最近提出的 RADF (Regional Attention-based Deep Feature)[35]，RADF 同时使用全局信息和局部信息来生成区域特征的权值. MSBD 在 Oxford5K 数据集上甚至取得了与 CIR(CNN Image Retrieval)[32]相同的 mAP. CIR 的 CNN 是一个二分支的孪生网络，而且 CIR 使用了额外的大数据集来进行训练，并使用了困难负例挖掘(hardnega-tive mining)方法. MSBD 是一个二阶段的方法，CIR 是一个端到端的方法. MSBD 的 mAP 在 Oxford5K 上不如 DIR[31]，DIR 的框架是一个三分支网络，而且使用了额外的 Landmark 数据集来进行训练. 此外，DIR 还使用了区域生成网络来生成感兴趣的区域，而 MSBD 的浮点描述符使用的是 EdgeBox. 尽管 MSBD 没有在 Oxford5K 数据集上取得最好的结果，但是，MSBD (D=256)的大小仅为 32 个字节，而 DIR 的大小是 2048 个字节，是 MSBD 的 64 倍. MSBD 与浮点描述符在 Paris6K 数据集上的比较如表 6 所示. 从表 6 可以看出，MSBD 在 Paris6K 数据集上仅用 256 位就超过了所有的方法，MSBD 的准确率比 CIR, DIR 和 RADF 分别要高出 6.8%、4.7%和 3.5%. 从表 5 和表 6 可以看出，MSBD(binary)的 mAP 甚至要高于浮点的 MSBD-float，这是因为浮点描述符 MSBD-float 使用的是 L2-欧氏距离，而 MSBD(binary)使用的是本文提出的不相似性度量，这表明了本文不相似性度量的有效性. 在大数据集 Oxford105K 上，为了适应这个数据集的大小采用了不同的设置：在全局分支，将图片缩放到 513，而不是 1024；在对象分支，仅用一个尺度 1.0；直接用汉明距离计算不相似性度量，因为在 100K flickr 图片上计算的类概率向量不准确. 在训练集中加入了 1000 张 flickr 图片. 从表 5 可以看出，准确率受到了一定程度的影响，但 MSBD 的准确率依然超过了 R-MAC[6] (float)，CroW[5](float)，并且 MSBD(D=2048)以 2048 位的空间不仅取得了与 MSBD-float 几乎一致的准确率，而且还与最近提出的 RADF 的准确率比较相近，而 RADF 的大小是 MSBD 的 32 倍.

表 6 MSBD 与浮点描述符在 Paris6K 数据集上的比较

	D	Paris6K
R-MAC[6] (float)	512	83.0
CroW[5] (float)	512	79.7
Faster-RCNN[70](float)	4096	79.8
NetVLAD[33](float)	256	73.5
RADF[35](float)	2048	87.5
CIR[32](float)	512	85.0
DIR[31](float)	512	87.1
MSBD-float(D1=512)	2048	89.4
MSBD(binary)	256	**91.8**

在 Holidays 数据集上，MSBD-float 的各个分支的描述符维度 $D1$ 被设为 256，而且全局分支还融合了全连接层特征，这个全连接层位于 *softmax* 层之前. 先将图片分为 1×2，然后分别提取这两个分片以及整幅图片的全连接层特征，将它们分别 2-范数标准化，用 PCA 降到 256 维，再次 2-范数标准化，然后串连起来，再将这个串连起来的特征与全局分支的描述符相串连. 最后得到的 MSBD-float 的维度是 1792 维. MSBD 与浮点描述符在 Holidays 数据集上的比较如表 7 所示. 从表 7 可以看出 MSBD 仅用 256 位就超过了所有其它的方法. 在 512 位（64 个字节）时，MSBD 的准确率比 CIR(float)，DIR(float)分别要高出 6.3%，2.1%，证明了 MSBD 的有效性.

表 7 MSBD 与浮点描述符在 Holidays 数据集上的比较

	D	Holidays
TE+DA[72]	8064	77.1
FAemb[73]	15525	78.7
Nerual codes[1](float)	4096	79.3
MOP_CNN[3](float)	2048	80.2
SpoC[4](float)	256	80.2
CroW[5](float)	512	85.1
NetVLAD[33](float)	256	82.1
CIR[32](float)	512	82.5
DIR[31](float)	512	86.7
MSBD-float(D1=512)	1792	**91.4**
MSBD(binary)	512	**88.8**

5 结 论

本文提出了一个占用空间小、区分力强的图像描述符-多层语义二值描述符(Multi-level Semantic Binary Descriptor, MSBD)，以用于图像检索. MSBD 同时融入了全局信息和局部信息，具有很强的区分力，在图像检索任务上取得了优良的性能. 为了得到一个占空间小且拥有较高区分力的二值图像表示，结合稀疏性提出了一个迭代的二值描述符学习算法. 此外，还提出了一个不相似性度量以融合哈希不相似度以及类级别的不相似性，可以有效提高

查询的准确率. 尽管 MSBD 是一个二阶段的方法，MSBD 在 Oxford5K, Paris6K 以及 Holidays 数据集上的准确率不仅超过了传统的哈希方法，还超过了同类中的非监督式和监督式端到端深度哈希方法，证明了其具有较强的区分力. MSBD 在小一个量级的情况下，甚至仅用 32 个字节就超过了很多当前最先进的浮点描述符方法，证明了 MSBD 的有效性.

参 考 文 献

[1] Artem Babenko, Anton Slesarev, Alexander Chigorin, Victor S. Lempitsky. Neural codes for image retrieval//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 584-599

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

[3] Yunchao Gong, Liwei Wang, Ruiqi Guo, Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 392-407

[4] Yandex, Artem Babenko, Lempitsky, Victor. Aggregating local deep features for image retrieval//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2016: 1269-1277

[5] Yannis Kalantidis, Clayton Mellina, Simon Osindero. Cross-dimensional weighting for aggregated deep convolutional features//Proceedings of the European Conference on Computer Vision Workshops. Amsterdam, The Netherlands, 2016: 685-701

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

[7] Herve Jegou, Florent Perronnin, Matthijs Douze et al. Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(9): 1704-1716

[8] Josef Sivic, Andrew Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos//Proceedings of the IEEE International Conference on Computer Vision. Nice, France, 2003: 1470-1477

[9] Jorge Sanchez, Florent Perronnin, Thomas Mensink, Jakob J. Verbeek. Image classification with the fisher vector: Theory and Practice. International Journal of Computer Vision, 2013, 105(3): 222-245

[10] Mayur Datar, Nicole Immorlica, Piotr Indyk, Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions// Proceedings of the 20th ACM Symposium on Computational Geometry. New York, USA, 2004: 253-262

[11] Yair Weiss, Antonio Torralba, Robert Fergus. Spectral hashing// Proceedings of the Advances in Neural Information Processing Systems. British Columbia, Canada, 2008: 1753-1760

[12] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2916-2929

[13] Brian Kulis, Trevor Darrell. Learning to hash with binary reconstructive embeddings//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, British Columbia, 2009: 1042-1050

[14] Kaiming He, Fang Wen, Jian Sun. K-means hashing: An affinity-preserving quantization method for learning binary compact codes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2938- 2945

[15] Wang J, Shen H T, Song J, et al. Hashing for similarity search: A survey. arXiv preprint arXiv:1408.2927, 2014

[16] Wang J, Zhang T, Sebe N, et al. A survey on learning to hash. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 769-790

[17] Do T T, Doan A D, Cheung N M. Learning to hash with binary deep neural network//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 219-234

[18] Lai H, Pan Y, Liu Y, et al. Simultaneous feature learning and hash coding with deep neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3270-3278

[19] Shen F, Shen C, Liu W, et al. Supervised discrete hashing// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 37-45

[20] Liu Hao-Miao, Wang Rui-Ping, Shan Shi-Guang, Chen Xilin. Learning to hash with discrete optimization. Chinese Journal of Computers, 2019, 42(5): 1149-1160 (in Chinese)
(刘昊淼，王瑞平，山世光，陈熙霖. 基于离散优化的哈希编码学习方法. 计算机学报，2019，42(5):1149-1160)

[21] Lin K, Lu J, Chen C S, et al. Learning compact binary descriptors with unsupervised deep neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1183-1192

[22] Lin K, Lu J, Chen C S, et al. Unsupervised deep learning of compact binary descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(6): 1501-1514

[23] Liu H, Wang R, Shan S, et al. Deep supervised hashing for fast image retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2064-2072

[24] Li Q, Sun Z, He R, et al. Deep supervised discrete hashing//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 2482-2491

[25] Luo X, Wu Y, Xu X S. Scalable supervised discrete hashing for large-scale search//Proceedings of the 2018 World Wide Web Conference. Lyon, France, 2018: 1603-1612

[26] Kang W C, Li W J, Zhou Z H. Column sampling based discrete supervised hashing//Proceedings of the Thirtieth AAAI conference on Artificial Intelligence. Phoenix, USA, 2016: 1230-1236

[27] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, United States, 2012: 1097-1105

[28] Donahue J, Jia Y, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition//Proceedings of the International Conference on Machine Learning. Beijing,

China, 2014: 647-655

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

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

[31] Gordo A, Almazán J, Revaud J, et al. Deep image retrieval: Learning global representations for image search//Proceedings of the European Conference on Computer Vision.Amsterdam, The Netherlands, 2016: 241-257

[32] Radenovi F, Tolias G, Chum O. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 3-20

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

[34] Agarwal S, Furukawa Y, Snavely N, et al. Building rome in a day. Communications of the ACM, 2011, 54(10): 105-112

[35] Kim J, Yoon S E. Regional attention based deep feature for image retrieval//Proceedings of the British Machine Vision Conference. Newcastle, UK, 2018: 209

[36] Kulis B, Grauman K. Kernelized locality-sensitive hashing for scalable image search//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 2130-2137

[37] Cao Y, Zhang H, Guo J. Weakly supervised locality sensitive hashing for duplicate image retrieval//Proceedings of the 2011 18th IEEE International Conference on Image Processing. Brussels, Belgium, 2011: 2461-2464

[38] Jiang K, Que Q, Kulis B. Revisiting kernelized locality-sensitive hashing for improved large-scale image retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4933-4941

[39] Heo J P, Lee Y, He J, et al. Spherical hashing//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 2957-2964

[40] Li Z, Liu X, Wu J, et al. Adaptive binary quantization for fast nearest neighbor search//Proceedings of the Twenty-second European Conference on Artificial Intelligence. The Hague, The Netherlands, 2016: 64-72

[41] Xia Y, He K, Kohli P, et al. Sparse projections for high-dimensional binary codes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3332-3339

[42] Song J, He T, Gao L, et al. Deep region hashing for generic instance search from images//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 402-409

[43] Do T T, Hoang T, Le Tan D K, et al. Binary constrained deep hashing network for image retrieval without manual annotation// Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa Village, USA, 2019: 695-704

[44] Xia R, Pan Y, Lai H, et al. Supervised hashing for image retrieval via image representation learning//Proceedings of the Twenty-eighth AAAI conference on Artificial Intelligence. Quebec City, Canada, 2014: 2156-2162

[45] Zhao F, Huang Y, Wang L, et al. Deep semantic ranking based hashing for multi-label image retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1556-1564

[46] Song G, Tan X. Hierarchical deep hashing for image retrieval. Frontiers of Computer Science, 2017, 11(2): 253-265

[47] Ou X, Ling H, Liu S, et al. Hierarchical deep semantic hashing for fast image retrieval. Multimedia Tools and Applications, 2017, 76(20): 21281-21302

[48] Yang H F, Lin K, Chen C S. Supervised learning of semanti-cs-preserving hash via deep convolutional neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(2): 437-451

[49] Cao Z, Long M, Wang J, et al. Hashnet: Deep learning to hash by continuation//Proceedings of the IEEE International Conference on Computer Vision.Venice, Italy, 2017: 5608-5617

[50] Li Q, Sun Z, He R, et al. Deep supervised discrete hashing// Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 2482-2491

[51] Cao Y, Long M, Wang J, et al. Deep quantization network for efficient image retrieval//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 3457-3463

[52] Jegou H, Douze M, Schmid C. Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(1): 117-128

[53] Duan Y, Lu J, Wang Z, et al. Learning deep binary descriptor with multi-quantization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1183-1192

[54] Yu T, Yuan J, Fang C, et al. Product quantization network for fast image retrieval//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 186-201

[55] Klein B, Wolf L. End-to-end supervised product quantization for image search and retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5041-5050

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

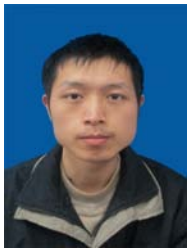
[57] Liu L, Wang L, Liu X. In defense of soft-assignment coding//Proceedings of the 2011 International Conference on Computer Vision. Barcelona, Spain, 2011: 2486-2493

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

[59] Zitnick C L, Dollá r P. Edge boxes: Locating object proposals from edges//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 391-405

[60] Pont-Tuset J, Arbelaez P, Barron J T, et al. Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(1): 128-140

[61] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 91-99
[62] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 379-387
[63] Hou Q, Cheng M M, Hu X, et al. Deeply supervised salient object detection with short connections//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3203-3212
[64] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014
[65] Xie S, Tu Z. Holistically-nested edge detection//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1395-1403
[66] Arandjelovi R, Zisserman A. Three things everyone should know to improve object retrieval//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 2911-2918
[67] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778
[68] Philbin J, Chum O, Isard M, et al. Lost in quantization: Improving particular object retrieval in large scale image databases// Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8
[69] Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching//Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8
[70] Salvador A, Giro-i-Nieto X, Marques F, et al. Faster R-CNN features for instance search//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, USA, 2016: 9-16
[71] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015
[72] Jégou H, Zisserman A. Triangulation embedding and democratic aggregation for image search//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 3310-3317
[73] Do T T, Tran Q D, Cheung N M. FAemb: A function approximation-based embedding method for image retrieval //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3556-3564



WU Ze-Bin, Ph.D. candidate. His research interests include image retrieval and deep learning.

YU Jun-Qing, Ph.D., professor. His research interests include digital media processing and multi-core programming environments.

HE Yun-Feng, Ph.D., associate professor His research interests include digital video processing and retrieval.

GUAN Tao, Ph.D., associate professor. His research interests include mobile visual search, augmented reality and computer vision.

Background

As the explosive growing of the multimedia data on the Internet, content-based image retrieval is attracting increasing attention. The performance of an image retrieval system is largely decided by the descriptor used. Many traditional shallow descriptor building frameworks have been proposed, however, the accuracy they achieve is not satisfying, especially for complex and large scale data-sets. With the advent of deep learning, making use of convolutional neural network to learn compact and discr-iminative representation has attracted considerable inter-est. However, most of the time, just the fully-connected layer or the last convolutional layer is used to build the descriptor. The features from the fully-connected layer usually capture the high-level semantic information and lack local information. To tackle this problem, we have proposed a multi-level semantic binary descriptor (MSBD) to build a compact and discriminative binary descriptor for image retrieval task, saving storage and preserving the retrieval accuracy at the same time. Our MSBD captures global and local information at the same time and is informative. An iterative learning strategy is proposed to learn discriminative binary descriptor without much loss of information. To further improve the retrieval accuracy, a dissimilarity function is proposed to fuse the hash-level dissimilarity and class-level dissimilarity. The experiments demonstrate the effectiveness of our method, our method even outperforms many state-of-the-art real-valued desc-riptors with far less bytes.

Before this work, our group has worked on the topic of image retrieval for more than five years. Some methods proposed by our group have been published in SCI journals. This work was supported in part by the National Natural Science Foundation of China (No. 61572211, 61173114, 61202300).