doi:10.3772/j.issn.1002-0470.2019.03.002

基于多层次特征表示的图像场景分类算法①

顾广华②*** 秦 芳③***

(*燕山大学信息科学与工程学院 秦皇岛 066004)

(**河北省信息传输与信号处理重点实验室 秦皇岛 066004)

摘 要 传统场景分类采用底层尺度不变特征变换(SIFT)特征,运用词袋(BoW)模型以及空间金字塔(SPM)模型进行分类判别。然而,单一的低层描述的识别精度有限,无法有效表征内容多变的场景图像。本文提出基于多层次特征表示的图像场景分类算法,利用滑动窗均匀采样图像块,分别提取图像块的密集SIFT 特征和卷积层卷积神经网络(CNN)特征,使用聚集局部描述符编码(VLAD)方法分别编码图像块的局部特征,将一幅图像的多个图像块特征顺序级联形成该幅图像的描述,由此构建包含局部语义信息的低层图像描述和中层图像描述。与此同时,将图像的低层描述与中层描述融合到图像的全连接层的高层语义中,从而获得整合了局部空间信息和全局语义信息的精确图像表示。本文在两个常用的场景数据集上进行了分类实验,结果表明,融合多层次特征描述的图像表示能够取得更好的分类结果。

关键词 低层描述,中层描述,高层语义,聚集局部描述符编码(VLAD)编码,场景分类

0 引 言

场景图像分类在计算机视觉和模式识别等领域一直是非常热门的研究方向,旨在通过对图像进行特征描述,将场景内容相似的图像归为一类。场景图像分类有广泛的应用前景,如图像检索、智能机器人、无人驾驶等。然而,场景图像分类面临着许多挑战,不仅存在大的内类差异和类间相似性,而且场景图像中常常存在复杂多变的目标内容。

场景图像分类任务中,高效的特征表示是实现精确分类的前提。在早期的研究中,一方面有利用主动学习技术进行建模来获得有效信息的方法[1,2];另一方面,采用全局特征来表示图像,如颜色、纹理、形状信息等。这些特征缺乏语义信息,且对尺度变换和光照遮挡缺乏鲁棒性,难以处理高度变化的复杂场景。为解决这个问题,Lowe 等人[3]提出尺度不变特征变换(scale invariant feature transform,SIFT)特征,该特征在图像尺度、方向等因素改变时具有不变性,在光照、遮挡和三维视角改变时具有较强的鲁棒性,能够有效地降低噪声等因素的影响[4-6]。自此,一系列基于 SIFT 特征的图像分类算法被提出,其中词袋(bag of words, BoW)[3]模型赢得了巨大的关注和广泛的应用,尤其是在图像检索和图像分类领域。然而,BoW 表示不具有足够的描述能力,它是分配给每个视觉单词的图像描述符数量的统计直方图,通过硬划分的方式来形成图像表示,丢失了图像的空间信息,从而影响了分类准确性。为了解决这个问题,空间金字塔匹配(spatial pyramid matching, SPM)[7]模型被提出,它将图像划

① 国家自然科学基金(61303128),河北省自然科学基金(F2017203169, F2018203239),河北省高等学校科学研究重点项目(ZD2017080)和河北省留学回国人员科技活动(CL201621)资助项目。

② 男,1979 年生,博士,教授;研究方向:图像场景分类;E-mail: guguanghua@ysu.edu.cn

③ 通信作者,E-mail: qinfang940506@163.com

(收稿日期:2018-06-15)

分成子区域,统计每个子区域里的直方图,再将所有子区域的直方图级联来形成图像描述,增加了图像的空间信息,一定程度上提高了分类精度。然而这两种图像描述都是通过特征描述子计数形成的,损失了特征所包含的图像信息。因此,研究者们提出了一系列新的特征编码算法,如局部约束线性编码(locality constrained linear coding, LLC)[8]、Fisher 矢量编码(Fisher vector, FV)[9]、聚集局部描述符编码(vector of locally aggregated descriptors, VLAD)[10]等。LLC 编码利用局部约束将每个描述子投影到它的局部坐标系中,并且投影坐标通过特征各维最大池化整合来产生最终的图像表示。FV 本质上是用似然函数的梯度向量来表达一幅图像,它结合了生成式方法和判别式方法的优势,采用高斯混合模型(Gaussian mixture model, GMM)来估计图像描述的特征分布,但是它基于使用高斯混合模型的视觉词典学习方法,不足以保证字典的判别性,且计算耗时。VLAD 是 FV 的非概率版本,它使用 K 均值聚类(K-means)代替高斯混合模型聚类,计算局部特征与其最近邻视觉词之间的累积残差,更充分地利用图像的局部特征,计算代价相对较小,已被成功应用于图像分类和检索任务[11-13]。

传统的基于 VLAD 编码的图像分类方法通常使用 SIFT 特征,且取得了良好的分类性能,然而近年来基于深度学习的方法在分类准确性上取得了飞跃性的突破。尤其是卷积神经网络(convolutional neural networks, CNN)[14]显著提高了各种视觉任务的性能,CNN 可以学习更鲁棒和丰富的中层图像描述符。因此,考虑到算法的效率和性能,以及不同特征之间的优势互补,本文提出基于多层次特征表示的图像场景分类算法,利用滑动窗将图像划分为大小相同的图像块,提取图像块的密集 SIFT 特征,使用 VLAD 方法编码图像块的局部特征,将一幅图像的多个图像块特征顺序级联形成该幅图像的描述,将此描述作为图像的低层描述。与此同时,提取图像块的卷积层 CNN 特征,同样进行 VLAD 编码,将图像块特征级联作为图像的中层描述。最后,将低层描述与中层描述融合到图像的全连接层的高层语义中,从而获得含有丰富语义信息的精确图像表示。

1 多层次特征表示

SIFT 特征自提出以来,被广泛用于图像处理的各个领域,是一种鲁棒性强的局部特征,对旋转、尺度缩放、亮度变化保持不变性,对视角变化、仿射变换、噪声也保持一定程度的稳定性。相反,CNN 特征是一种全局图像表示,可以用来识别位移、缩放及其他形式扭曲不变性的二维或三维图像,在视觉识别任务尤其是图像分类领域取得了巨大的成功[15-17]。CNN 是一种带有卷积结构的深度神经网络,通常至少有 2 个非线性可训练的卷积层,2 个非线性的固定卷积层(又叫池化层)和 1 个全连接层,一共至少 5 个隐含层。它采用卷积层与池化层交替设置,通过卷积层提取出特征,再进行组合形成更抽象的全局描述。因此,本文利用局部图像块的 SIFT 特征构建图像的低层描述,利用图像块的卷积层 CNN 特征构建中层图像描述,获取原始图像在第一个全连接层的特征,作为高层语义特征,将低层描述与中层描述整合到图像的高层语义中去,获得更精确的图像表示;最后,采用线性支持向量机(support vector machine, SVM)来实现图像的分类判别。其原理框图如图 1 所示。

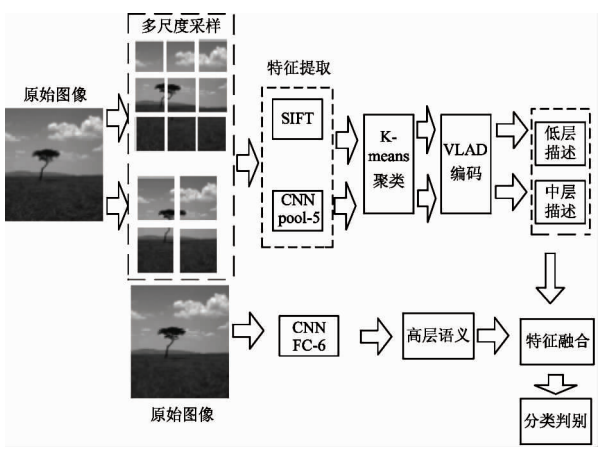

图 1 多层次特征表示的图像场景分类算法原理框图

该模型一方面对原始图像进行多尺度采样,提取多尺度图像块的特征描述子,进行 VLAD 编码,生成的图像描述相比单尺度图像的特征含有更多的局部空间信息;另一方面,使用 SIFT 和 CNN 这两种特

征来表征图像,充分结合 SIFT 特征对于局部信息刻画能力的优势和 CNN 特征对于图像语义信息表达的准确性,使其相互补充;此外,考虑到场景图像的特殊性,即场景图像的全局结构信息对于分类识别至关重要,因此,为了保留图像的全局信息,对原始图像提取其全连接层的语义特征,将三个层次的特征进行融合,从而获得更加准确的图像描述。

1.1 图像块特征提取

为了获得包含更多局部语义信息的图像描述,本文通过对原始图像均匀采样,使用所有图像块顺序级联的特征来表征该幅图像,不仅可以加强图像的局部信息表达,也能增加图像的空间结构信息。具体而言,首先将图像调整为 256×256 大小并转化为灰度图,将其用作原始图像。其次,使用长度为 128 像素的滑动窗分别以 128 像素和 64 像素步长对图像进行采样,获得尺度均为 128×128 图像块。然后,提取图像块的密集 SIFT 特征,与此同时,使用预训练的卷积神经网络 VGG-F 模型提取图像块在最后一个卷积层的 CNN 特征。因为卷积层特征主要含有图像的局部语义信息,且特征维度较低;而全连接层的特征主要表征图像的全局语义信息,特征维度较高,本文为了获得含有更多局部语义信息的中层图像描述,同时减小计算代价,故采用卷积层特征,而非通用的全连接层特征。其中,VGG-F 网络架构由 5 个卷积层和 3 个全连接层组成,卷积层和池化层交替排列,如图 2 所示。

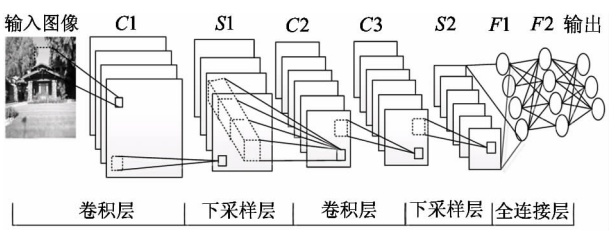

图 2 预训练的卷积神经网络架构示意图

图 2 中 C 表示卷积层,S 表示池化层,F 表示全连接层。卷积层通过图像与滤波器的卷积运算,使原信号特征增强,并且噪音降低。池化层通过局部非线性运算来减小输入层的空间尺寸,减少计算量的同时保持图像的旋转不变性。全连接层采用 softmax 全连接,得到的激活值即卷积神经网络提取到的特征。

1.2 VLAD 编码

获得局部图像块的密集 SIFT 特征和深度卷积层特征之后,若图像块特征为 D 维向量,对于训练集的图像块特征使用 K-means 聚类算法来生成具有 M 个聚类簇的通用视觉码本,基于获得的视觉码本 B,采用 VLAD 编码来获得图像的视觉描述符。具体而言,首先把每个描述子赋给离它最近的码本,求出残差向量,即所有特征向量与其类中心向量的差:

$$NN(\boldsymbol{x}_t) = \arg\min_{b_i} \| \boldsymbol{x}_t - \boldsymbol{b}_i \| \tag{1}$$

其次,将同类别的残差向量求和,得到 M 个 D 维的向量 $\boldsymbol{v}_i$:

$$\boldsymbol{v}_i = \sum_{x_t: NN(x_t) = b_i} \boldsymbol{x}_t - \boldsymbol{b}_i \tag{2}$$

最后,将所有区域得到的向量串联在一起并做 L_2 归一化,得到大小为 $\boldsymbol{K} = M \times D$ 的一维编码向量。

$$\boldsymbol{v} = [v^1, v^2, \cdots, v^M] = [v_1, v_2, \cdots, v_K] \tag{3}$$

L_2 范数归一化的目的主要是为了使得特征向量范数为 1,使得对特征的比较是在同一个尺度上,比如可以用来减少同一个物体在不同光照下由于光照等因素带来的特征差异。

2 多层次特征融合

对训练集和测试集图像块的密集 SIFT 特征和卷积层 CNN 特征进行 VLAD 编码之后,将每一幅图像的所有图像块的编码特征顺序级联作为该幅图像的图像描述。相应地,针对图像块的 SIFT 编码特征和 CNN 编码特征构建图像的低层和中层特征描述。由于构建的低层与中层图像描述仍然仅仅表征图像的局部语义信息,缺乏图像的全局描述。对于场景图像而言,图像的全局结构信息对于分类识别至关重要。因此,为了整合场景图像的全局信息,本文提取原始图像在第一个全连接层的全局 CNN 特征。因为图像经卷积和池化层抽象后,全连接层输出的特征丢失了目标的详细信息和场景类别的空间信息,主要表征图像的高层语义信息。由于本文提取的特征不同,其特征维度也各不相同,其中密集

SIFT 特征为 128 维,卷积层 CNN 特征为 256 维,第一个全连接层的全局 CNN 特征为 4 096 维。当使用 $M=200$ 的码本进行 VLAD 编码,获得的低层图像描述和中层图像描述的特征维度分别为 $1\times12\,800$ 和 $1\times51\,200$。为了实现全局信息与局部信息的融合,获得图像的高效表示,本文将获得的低层图像描述和中层图像描述与图像的全局 CNN 特征级联,使得最终的图像表示包含更加丰富的语义信息。由于不同层次的特征维度相差较大,为了平衡各层次特征对于最终图像描述的影响,同时考虑到不同图像描述包含的图像信息量不同,对 3 种层次的图像描述进行加权级联,即:

$$f_{\text{final}}=[w_1\times f_1,\ w_2\times f_2,\ w_3\times f_3] \tag{4}$$

其中,f_1 表示低层图像描述,f_2 表示中层图像描述,f_3 表示全局 CNN 特征。由于图像块的 CNN 特征比密集 SIFT 特征包含更多的语义信息,对于分类判别的作用更大,因此,本文以 $w_1=0.2$,$w_2=0.4$,$w_3=0.4$ 的权重比例对 3 种特征描述进行加权级联,获得具有丰富语义信息的融合特征表示。

3 实验结果与分析

3.1 数据集

本文提出的场景分类方法是针对 2 个通用的场景数据集进行评估的:15 类场景数据集 15-category 和 SUN397 数据集。其中 15-category 数据集中的每个类别包含 210 到 410 幅图像,总共有 4 486 张灰度值图像。图像的平均大小约为 300×250。另一个是 SUN397 场景识别数据集[18]。它包含 397 个场景类别,大约 10 万张图像,每类至少 100 张图像,是目前最大的场景数据集。为了验证本文所提方法的有效性,并减少时间耗费,本文从 SUN397 中选择 15 个具有挑战性和代表性的类别来组成一个新的数据集 SUN397-15,包括“ball _ pit”,“wave”,“bull-ring”,“rock _ arch”,“subway _ interior”,“ice _ skating _ rink _ indoor”,“sky”,“bamboo _ forest”,“bow _ window _ outdoor”,“pagoda”,“skatepark”,“electrical _ substation”,“ocean”,“shower”,“train _ station _ platform”。其中既有人造室内场景,又有人造室外场景,还有自然场景,且还含有“ocean”和“sky”这类十分容易混淆的类,因此在该数据集上的实验结果能够充分展现整体的分类性能。实验中,数据集的每个类别被随机地划分到训练集和测试集中。每次选择 80 幅图像用于训练,20 幅图像用于测试。实验重复 10 次不同的划分,取 10 次实验的平均识别率作为最终结果。

3.2 码本大小的选择

K-means 聚类的缺点之一是对 M 值的选择敏感,因此本文选择 $M=50$、100、150、200、250 这 5 种不同大小的码本,从中找出最佳尺寸。首先使用长度为 128 像素的滑动窗以 128 像素步长对图像进行采样,提取图像块的 CNN 特征,执行 K-means 聚类,获得不同尺寸的通用视觉码本,基于通用视觉码本,使用 VLAD 编码方法对局部图像块的 CNN 特征进行编码,将一幅图像的所有图像块的编码特征级联,形成图像的中层描述,并进一步将其与原始图像的第一个全连接层的全局 CNN 特征等比例加权级联,在 15-category 数据集上比较了分类准确性,分类结果如表 1 所示,表 1 中的编码特征(CNN)即图像的中层描述。与此同时,为了验证码本大小对于不同特征的泛化能力,本实验进一步使用 SIFT 特征来求得最佳尺寸的码本,由于图像的密集 SIFT 特征数量较多,故此时不对图像进行滑动窗采样,直接提取原始图像的密集 SIFT 特征进行 VLAD 编码,构建低层图像描述,同样将其与全连接层的全局特征进行级联,在 15-category 数据集上比较了分类准确性,分类结果如表 2 所示,表 2 中的编码特征(SIFT)即图像的低层描述。

表 1 使用 CNN 特征时不同尺寸码本的分类比较

类码本	编码特征(CNN)	融合特征
$M=50$	85.67%	91.00%
$M=100$	87.67%	91.33%
$M=150$	88.00%	91.67%
$M=200$	88.67%	92.00%
$M=250$	86.67%	90.67%

表 2　使用 SIFT 特征时不同尺寸码本的分类比较

类码本	编码特征(SIFT)	融合特征
$M=50$	64.67%	89.00%
$M=100$	65.33%	90.00%
$M=150$	70.00%	90.00%
$M=200$	74.33%	91.00%
$M=250$	72.67%	90.33%

为了更加直观地看出使用不同特征获得的编码特征与级联特征的分类准确性随码本尺寸的变化情况,将表 1、2 绘制成柱状图,如图 3 所示。

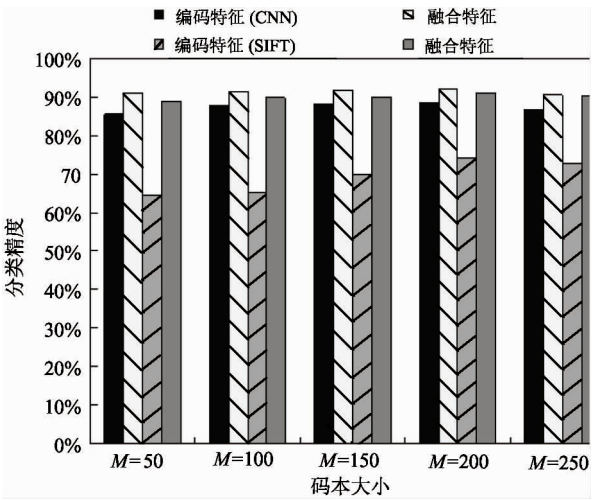

图 3　不同特征的分类精度随码本大小的变化

由图 3 可以看出,随着码本尺寸的增加,使用密集 SIFT 特征获得的编码特征与使用采样图像块的卷积层 CNN 特征获得的编码特征的分类精度变化趋势均为先增加后减小,且当码本尺寸为 $M=200$ 时,均能够获得最佳的分类准确性。与此同时,分别将低层和中层编码特征与高层语义特征级联获得的融合特征在数据集上的分类精度遵循同样的规律。因此,在本文实验中,均采用尺寸为 $M=200$ 的码本。

3.3　多尺度对比

为了验证本文方法的泛化能力,进行了进一步的扩展实验,除了使用长度为 128 像素的滑动窗以 128 像素步长对图像进行采样之外,选择相同尺度的滑动窗以不同步长进行局部图像块采样。具体而言,使用长度为 128 像素的滑动窗以 64 像素步长对图像进行采样,分别对获得的不同数量的图像块提取密集 SIFT 特征和卷积层 CNN 特征,以 $M=200$ 的码本大小执行聚类,使用 VLAD 方法进行特征编码,构建不同尺寸的低层和中层图像描述,进一步分别将其与原图的全连接层全局特征融合,在 15-category 和 SUN397-15 两个数据集上比较了分类准确性,分类结果如表 3、4 所示。同时,为了表明该方法能够更好地刻画图像的空间信息,本文也对原始图像直接提取密集 SIFT 特征和卷积层 CNN 特征,不对图像进行滑动窗采样,构建多层次特征表示。表 3、4 中 Scale1 表示不采样图像块,直接对原始图像的特征构建多层次图像描述,Scale2 表示以步长为 128 像素采样 128×128 的图像块构建多层次图像描述,Scale3 表示以步长为 64 像素采样 128×128 的图像块构建多层次图像描述,并进一步将不同层次的图像描述与原始图像的第一个全连接层的 CNN 特征级联,获得更加有效的融合图像表示。

表 3　15-category 上多尺度编码特征与融合特征的分类比较

15-category	Scale1(%)	Scale2(%)	Scale3(%)
低层描述	73.43	76.37	82.03
低层+高层描述	90.97	91.30	92.20
中层描述	83.53	87.67	88.20
中层+高层描述	90.80	91.83	92.30

表 4　SUN397-15 上多尺度编码特征与融合特征的分类比较

SUN397-15	Scale1(%)	Scale2(%)	Scale3(%)
低层描述	82.50	87.57	90.70
低层+高层描述	97.17	97.20	97.43
中层描述	92.93	93.17	93.47
中层+高层描述	97.47	97.51	97.57

横向来看,表 3 中 Scale2 的低层描述的分类精度相比 Scale1 提高了 2.94%,中层描述的分类精度相比 Scale1 提高了 4.14%;Scale3 低层描述的分类精度相比 Scale1 提高了 8.60%,相比 Scale2 提高了 5.66%,中层描述的分类精度相比 Scale1 提高了 4.67%,相比 Scale2 也有所提高。由此可知,对图像进行滑动窗采样,将每个单独图像块的编码特征级联来构建图像描述,能够更加准确地刻画图像的

局部特征,且划分的块适当多的情况下,能够使图像描述包含更多的语义信息,因此,本文在构建多层次特征表示时选择长度为 128 像素的滑动窗以 64 像素步长对图像进行采样。纵向来看,在 Scale3 时中层描述的分类精度相比低层描述提高了 6.17%,将图像的低层和中层描述分别与图像在全连接层的全局特征融合后,低层 + 高层描述的分类精度相比低层描述提高了 10.17%,相比中层描述提高了 4.00%,中层 + 高层描述的分类精度相比中层描述提高了 4.10%,与低层 + 高层描述的分类精度相当。由此可以看出,使用局部图像块的 CNN 特征构建中层描述比使用 SIFT 特征构建的低层描述含有更多语义信息,且仅仅使用图像的局部特征构建图像描述不能很好地表征图像。对于场景图像分类问题,图像的全局特征至关重要,因此当图像的低层和中层描述分别与全局 CNN 特征融合后,均能取得令人满意的结果,且使用 SIFT 特征构建的低层描述获得的融合特征也能达到使用 CNN 特征的分类效果。由表 4 可知,在 SUN397-15 数据集上的分类精度与表 3 中 15-category 数据集遵循相同的规律。

3.4 多层次特征表示

本文使用滑动窗均匀采样图像块,提取其密集 SIFT 特征和卷积层 CNN 特征,构建多层次的图像描述,并进一步将不同层次的图像描述与图像的全局 CNN 特征融合,获得丰富的多层次特征表示。由前文的实验结果可知,采样适当多数量的图像块能够获得更多的局部语义信息,但如果图像块数量过多,会造成信息的冗余,反而影响分类精度。因此,选择以 64 像素步长对图像采样,即由一副图像获得 9 块尺度为 128 × 128 的图像块。依次构建图像的低层描述和中层描述,并提取图像的全局 CNN 特征。分别将不同层次的图像描述两两等比例加权级联,进行分类判别,并将 3 种层次的特征以 1:2:2 的权重比例加权融合,获得最终的图像表示。同时,为了更好地进行比较,本实验对由原始图像提取特征构建的多层次图像描述也进行了分类判别。在 15-category 和 SUN397-15 两个数据集上比较了分类精度,结果如表 5 所示。表 5 中 15-category-1 和 SUN397-15-1 表示直接对原始图像提取特征构建多层次特征表示,15-category-2 和 SUN397-15-2 表示对原始图像均匀采样构建多层次特征表示,且其中的高层描述均为原始图像的全连接层 CNN 特征,未经采样,此两种情况下相同。

表 5 两个场景数据集上多层次特征表示的分类比较

	15-category-1(%)	15-category-2(%)	SUN397-15-1(%)	SUN397-15-2(%)
低层描述	73.43	82.03	82.50	90.70
中层描述	83.53	88.20	92.93	93.47
高层描述	88.40	88.40	96.5	96.5
低层 + 中层描述	84.60	92.33	93.43	96.00
低层 + 高层描述	90.97	92.20	97.17	97.43
中层 + 高层描述	90.80	92.30	97.47	97.57
低层 + 中层描述 + 高层描述	91.40	93.33	97.33	98.00

由表 5 可以看出,两个数据集上单独层次的图像描述的分类精度为高层描述效果最好,中层描述次之,低层描述最低;当两种层次的图像描述加权级联后,低层 + 高层描述的分类精度与中层 + 高层描述相当,两者均高于低层 + 中层描述的分类精度,由此可知全局图像信息对于场景图像分类判别具有十分重要的作用。当 3 种层次的图像特征以不同比例加权融合后,能够取得最好的分类准确性。与此同时,对图像块均匀采样构建的多层次图像表示的分类效果均高于未经采样的图像特征,在 15-category 数据集上的最终融合特征的分类精度相比未采样的融合特征提高了约 2%,达到了 93.33% 的分类准确性。在 SUN397-15 数据集上的最终融合特征的分类精度相比未采样的融合特征提高了约 1%,达到了

98%的分类准确性。进一步验证了对图像进行适当采样,能够提高特征的表达能力。此处,给出本文最终采用低层 + 中层描述 + 高层描述的方法在两个数据集上分类结果的混淆矩阵,如图 4 和图 5 所示。

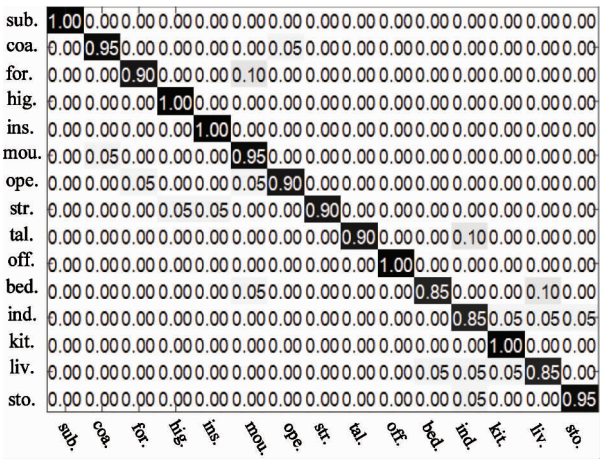

图 4　15 – category 数据集上分类结果的混淆矩阵

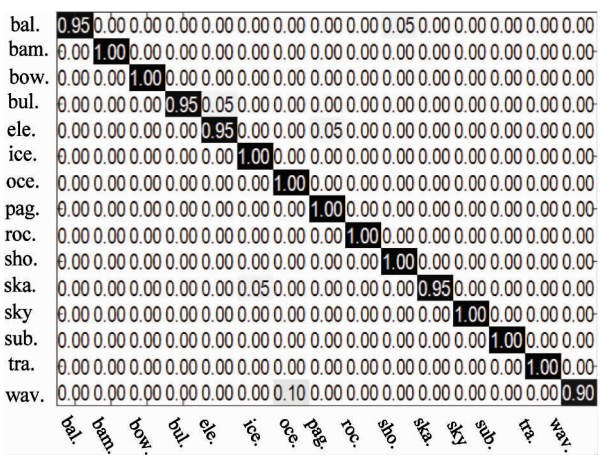

图 5　SUN397-15 数据集上分类结果的混淆矩阵

同时,为了显示本文提出的多层次特征表示分类算法的优越性,针对 15-category 数据集的识别,将本文的分类算法与其他分类方法[7,19-28]进行了对比,结果如表 6 所示。

由表 6 可知,本文的方法相比文献[21]中使用 VLAD 方法编码 SIFT 特征进行判别的分类精度提高了 15.98%,相比文献[22]中使用改进的 VLAD 方法编码 SIFT 特征进行判别的分类精度提高了 14.10%,相比文献[23]中使用改进的 VLAD 方法编码密集采样图像块的CNN特征的分类精度提高了 9.83%,相比文献[28]中使用深度 caffe 模型的方法在分类精度上也有所提高,且本文算法及结构的复杂性更低,运算成本更小,由此表明了本文方法的有效性。

表 6　15-category 数据集上不同分类方法比较

方法	15-category
BOW[19]	65.87%
GIST[20]	73.28%
VLAD[21]	77.35%
TNNVLAD[22]	79.23%
SPM[7]	81.40%
VLAD + CNN[23]	83.50%
Caffe[24]	87.99%
IFV[25]	89.20%
LScSPM [26]	89.80%
ISPR + IFV[27]	91.00%
DDSFL + Caffe[28]	92.81%
本文	93.33%

4　结 论

本文提出基于多层次特征表示的图像场景分类算法,利用滑动窗均匀采样图像块,分别提取图像块的密集 SIFT 特征和卷积层 CNN 特征,使用 VLAD 方法分别编码图像块的局部特征,将一幅图像的多个图像块特征顺序级联形成该幅图像的描述,由此构建包含局部语义信息的低层图像描述和中层图像描述。与此同时,将图像的低层描述与中层描述融合到图像的全连接层的高层语义中,从而获得整合了局部空间信息和全局语义信息的精确图像表示。在两个典型数据集上的实验结果,表明了本文所提出的场景分类方法的优越性。本文在将 3 个层次的特征描述进行融合时,直接采用加权级联的方式获得最终的图像表示,没有考虑各层次图像描述之间的关系,以及由于不同特征的维度差异过大而影响其在分类判别时的作用,且最终级联特征的维度过大,计算成本高。因此,下一步的工作是改进算法,对不同特征的维度进行调整,找到更好的特征融合方法,实现更好的分类判别。

参考文献

[1] Zhang X Y, Wang S, Yun X. Bidirectional active learning: a two-way exploration into unlabeled and labeled dataset[J]. *IEEE Transactions on Neural Networks and Learning Systems* (*TNNLS*), 2015, 26(12): 3034-3044

[2] Zhang X Y, Wang S, Zhu X, et al. Update vs. upgrade: modeling with indeterminate multi-class active learning [J]. *Neurocomputing* (*NEUCOM*), 2015, 162: 163-170

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

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

[5] Puggal S, Jindal S. Enhanced fingernail recognition based on GLCM, SIFT and NN[J]. *International Journal of Computer Applications*, 2018, 180(26):18-22

[6] Castillo-Carrión S, Guerrero-Ginel J E. SIFT optimization and automation for matching images from multiple temporal sources[J]. *International Journal of Applied Earth Observations & Geoinformation*, 2017, 57:113-122

[7] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, 2006. 2169-2178

[8] Wang J J, Yang J C, Yu K, et al. Locality-constrained linear coding for image classification[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA , 2012. 3360-3367

[9] Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization[C]. In: Proceedings of the Computer Vision and Pattern Recognition, Minneapolis, USA, 2007. 1-8

[10] Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation[C]. In: Proceedings of the Computer Vision and Pattern Recognition, San Francisco, USA, 2010. 3304-3311

[11] Jégou H, Chum O. Negative evidences and co-occurences in image retrieval: the benefit of PCA and whitening[C]. In: Proceedings of the European Conference on Computer Vision, Florence, Italy, 2012. 774- 787

[12] Arandjelovic R, Zisserman A. All about VLAD[C]. In: Proceedings of the Computer Vision and Pattern Recognition, Portland, USA, 2013. 1578- 1585

[13] Peng X, Wang L, Qiao Y, et al. Boosting VLAD with supervised dictionary learning and high-order statistics [C]. In: Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 2014. 660-674

[14] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]. In: Proceedings of the International Conference on Neural Information Processing Systems, Curran Associates Inc, USA, 2012. 1097-1105

[15] Cui Y Z, Cai Y H, Qiu C Y, et al. Scene detection of news video using CNN features[C]. In: Proceedings of the International Congress on Image and Signal Processing, Shanghai, China, 2018. 1-5

[16] Wang J, Luo C, Huang H Q, et al. Transferring pretrained deep CNNs for remote scene classification with general features learned from linear PCA network[J]. *Remote Sensing*, 2017,9(3):225-226

[17] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2016, 38(1):142-158

[18] Xiao J, Hays J, Ehinger K A, et al. SUN database: large-scale scene recognition from abbey to zoo[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA , 2010. 3485-3492

[19] Csurka G. Visual categorization with bags of keypoints [J]. *Workshop on Statistical Learning in Computer Vision ECCV*, 2004, 44(247):1-22

[20] Sampanes A C, Tseng P, Bridgeman B. The role of gist in scene recognition [J]. *Vision Research*, 2008, 48 (21):2275-2283

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

[22] Long X Z, Lu H T, Peng Y, et al. Image classification based on improved VLAD[J]. *Multimedia Tools & Applications*, 2016, 75(10):5533-5555

[23] Wang Q, Zhu J, Shao W, et al. Image classification based on deep local feature coding[C]. In: Proceedings

of the International Symposium on Intelligent Signal Processing and Communication Systems, Xiamen, China , 2017. 480-485

[24] Donahue J, Jia Y, Vinyals O, et al. DeCAF: A deep convolutional activation feature for generic visual recognition[C]. In: Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. 32: 647-655

[25] Vedaldi A, Fulkerson B. Vlfeat:an open and portable library of computer vision algorithms[C]. In: Proceedings of the International Conference on Multimedea 2010, Firenze, Italy, 2010. 1469-1472

[26] Gao S, Tsang W H, Chia L T. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2013, 35(1):92-104

[27] Lin D, Lu C W, Liao R J, et al. Learning important spatial pooling regions for scene classification[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA,2014. 3726-3733

[28] Zuo Z, Wang G, Shuai B, et al. Exemplar based deep discriminative and shareable feature learning for scene image classification [J]. *Pattern Recognition*, 2015, 48 (10):3004-3015

Image scene classification algorithm based on multi-level feature representation

Gu Guanghua[*,**], Qin Fang[*,**]

([*]School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)

([**]Hebei Key Laboratory of Information Transmission and Signal Processing, Qinhuangdao 066004)

Abstract

The traditional scene classification uses the bag of words (BoW) model and the spatial pyramid matching (SPM) model with scale invariant feature transform (SIFT) features for classification discrimination. However, the single low-level description fails to represent the scene images due to the complexity and variability of scenes. This paper proposes an image scene classification algorithm based on multi-level feature representations. The convolutional neural networks (CNN) features from the convolutional layer and the dense SIFT features of the image blocks, sampled by sliding windows evenly, are extracted and encoded by the vector of locally aggregated descriptors (VLDA) method, respectively. The encoding SIFT features and CNN features of multi-blocks are sequentially cascaded respectively to form a low-level description and middle-level description. Both descriptions contain the local semantic information of the image. Meanwhile, the low-level description and the middle-level description of the image are integrated into the high-level semantic features of the full-connected layer of the image, so that a more accurate image representation is obtained by integrating the local spatial information and the global semantic information. In this paper, scene classification experiments are performed on two commonly used scene datasets. The experimental results show that the fusion representation of multi-level feature descriptions achieves better classification results.

Key words: low-level description, middle-level description, high-level semantics, vector of locally aggregated descriptors(VLAD) coding, scene classification