图像场景分类技术综述

田艳玲[1,2,3],张维桐[1,2,3],张锲石[3],路 纲[1,2],吴晓军[1,2]
(1. 陕西师范大学现代教学技术教育部重点实验室,陕西西安 710062;2. 陕西师范大学计算机科学学院,陕西西安 710062;
3. 中国科学院深圳先进技术研究院,广东深圳 518055)

摘 要: 目前,基于计算机视觉分析的图像场景分类技术已被广泛研究并应用在众多学科领域中. 本文从不同角度对近年来典型的场景分类技术进行了深入的探讨与比较. 首先介绍了场景分类技术的背景、应用场景以及发展现状;然后基于特征提取、语义分析和机器学习的角度分别对国内外的相关研究进行系统的分析、比较及总结;最后对目前研究所面临的问题和未来技术的发展给出总结与展望.

关键词: 场景分类; 特征提取; 语义分析; 深度学习
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2019)04-0915-12
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.04.020

Review on Image Scene Classification Technology

TIAN Yan-ling[1,2,3], ZHANG Wei-tong[1,2,3], ZHANG Qie-shi[3], LU Gang[1,2], WU Xiao-jun[1,2]
(1. *Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi' an, Shaanxi* 710062, *China*;
2. *School of Computer Science, Shaanxi Normal University, Xi' an, Shaanxi* 710062, *China*;
3. *Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong* 518055, *China*)

Abstract: The computer vision based scene classification technology is widely developed and applied in different fields. In this paper, the typical scene classification technology is analyzed and compared from the different directions. First, the background, application and development situation is introduced. Then, the related researches both at home and overseas are analyzed, compared and summarized from the perspectives of feature extraction, semantic analysis and machine learning. Finally, the problems that the current researches are facing and potential future development are discussed.

Key words: scene classification; feature extraction; semantics analysis; deep learning

1 引言

图像场景分类技术涉及模式识别、计算机视觉系统、信号处理及人机交互等学科的交叉领域,是解决图片搜索、图像识别问题的关键技术,已成为计算机视觉领域中一个非常重要且极具挑战的研究课题. 早期,通过人工分类的方法还可以满足一些领域的基本需求,然而随着网络技术的普及及多媒体信息的爆炸性增长,图像内容种类和数量与日俱增. 如今面对海量的图像数据,依靠传统人工分类与标注的方式已经远远不能满足需求. 单纯依靠人工分类,不仅极大地浪费人力资源而且也不能保证工作的及时性与可靠性[1]. 如何让计算机以人类的思维逻辑去"理解"图像就是现在要着手解决的问题. 所谓图像场景分类是指对于已经给定的图像,通过判断识别它所包含的信息和内容来判断其所属的场景,从而达到分类的目的. 随着图像场景分类技术的不断发展,其内涵和范畴也在不断丰富拓展. 在图像分类发展早期,单幅图像的分类主要通过像素之间的关系权重来综合考虑,在人工干预下逐步做二类语义分割,确定场景类别. 之后,图像场景分类得到进一步发展,研究人员通过提取图像整体特征来确定该图像所属类别. 根据应用需求的不同,以及场景在组成架构、包含元素等方面的巨大差异,需要被分类的场景类别多达上千种. 目前公认的场景分类主要分为:自

收稿日期:2018-02-13;修回日期:2018-07-05;责任编辑:李勇锋
基金项目:国家重点研发计划(No. 2017YFB1402100);国家自然科学基金(No. 11772178, No. 61772508, No. U1713213);陕西省自然科学基金(No. 2017JM6101, No. 2017JM6103, No. 2017JM6060, No. 2017JQ6077);陕西师范大学中央高校基本科研业务费(No. GK201703060, No. GK201801004);陕西师范大学 2017 年度校级综合教改研究项目(No. 17JG33)

然场景、事件场景、城市场景和室内场景四大类[2]. 尽管图像场景分类研究已经取得很大的进步,但尚没有一种分类方法能在两类甚至多类图像场景数据库中兼具有良好的分类表现,这种矛盾在室内与室外的图像场景分类方法中表现的尤为明显.

在2006年美国麻省理工学院召开的场景理解研讨会上,图像场景分类技术第一次被明确定义为分类的一个关键课题. 自此图像场景分类技术在计算机视觉领域的重要地位基本奠定,每年都不断有大量新方法和贡献被提出. 在早期图像场景分类研究中,大多数方法是基于特征的,即通过描述颜色、纹理和形状等特征来实现分类[3,4]. 而后,用融合多种特征的方法来描述不同内容的图像场景,并较先前基于单层次特征的方法取得了更好的效果. 但由于无法进行深层涵义的理解,难以胜任复杂场景图像的分类. 而基于语义的场景分类方法由于层次化模型和分类器的发展得以推进[5,6],然而当场景种类需要被进一步细分时,上述方法就难以取得理想的分类效果. 深度学习的出现则通过大量数据学习到图像的共有深层特征来有效的解决这种问题.

2 场景分类方法

到目前为止,图像场景分类技术的发展主要经历了:基于特征、基于语义、基于学习三个阶段.

2.1 基于特征的分类方法

图像特征如颜色、纹理等是对图像特性的描述,可以描述图像的基本信息,也可以反映图像的深层结构信息. 下面就一些经典的特征分类方法进行分析与总结.

1999年David Lowe提出的图像局部特征描述算子(SIFT,Scale-Invariant Feature Transform)[7],由于其在尺度空间上的缩放、旋转和仿射变换不变性,在室外场景分类上精度较高,且能够较完整地描述图像的局部特征,当图像中尺度变化较大时也能准确识别图像的显著特征.

2001年由Aude Oliva等人模拟人的视觉提出GIST特征[3]以粗略提取图像以及其上下文信息. 该特征能够计算图像的整体特征,理解场景中的全局信息,并通过图像的能量谱信息提取图像的整体空间布局结构. 但随着图像内容及结构的复杂度不断增加,如分析粒度太过粗糙而忽略场景中物体的细节信息,导致分类结果远远偏离正确结果,增加了其复杂度和计算负担.

伴随着研究进展的推进,颜色、光线以及梯度特征在识别过程中表现出的不稳定性,成为当时图像场景分类问题的阻碍. 2005年提出的HOG(Histograms of Oriented Gradients)特征[8]延续了图像局部特征高识别度与准确的特点,采用梯度直方图方法有效地解决了光线、梯度特征等因素因敏感而引起局部场景轮廓识别率低下的问题. 然而HOG特征维度高、计算效率低,且具有很大冗余,没有考虑尺度变换对分类结果产生的影响. 为了改进HOG特征的不足,Anna Bosch通过使用金字塔模型在图像上分层抽取HOG特征并进行融合提出了PHOG(Pyramid HOG)特征[9]. 针对识别不同类别的场景和不同尺度的图像,研究者们还基于HOG提出了一系列改进的模型方法,如DPM(Deformable Part-based Models)[4]和之后基于DPM改进的SDPM(Supervised Object Localization with DPM)[10].

在随后的发展过程中,由于图像复杂度的提高,基于HOG特征的模型方法出现特征点冗余、计算效率低的问题. Jianxin Wu等人在2010年提出的CENTRIST特征[11]很好地解决了这一不足. 该特征通过对获取的像素点进行Census变换,并将其转化为统计直方图,形成CENTRIST特征来提取对象局部形状结构. 其转化后的图像依然保留整体和局部结构信息,因此能够模拟人类视觉系统的实时性,对物体的形状以及纹理等都有较明确的描述.

综上所述,四种特征被广泛使用,其优缺点如表1所示. 基于特征的场景分类算法通过提取图像特征(如颜色、形状和纹理等),结合特征描述并设计使用相应的分类器,在平衡了复杂度以及模型结构后取得了不错的场景分类效果.

表1 图像特征方法比较

名称	类型	输出	优点	缺点	适用场合	特点
GIST[3]	全局	光谱信息	计算复杂度低、简单易用	背景复杂、目标密集场景表现差	简单的自然场景	通过光谱表现整体布局
SIFT[7]	局部	邻域直方图	适用平移、旋转、尺度变换	复杂场景整体布局表现差	自然场景和简单的室内场景	通过梯度方向信息表现目标位置信息
HOG[8]		向量	表现轮廓和边缘	不稳定场景形状结构表现差	全局结构稳定的场景	通过像素块边缘梯度表现几何结构
CENTRIST[11]		Census变换值	表现局部特性、体现位置信息	复杂多变场景表现差	清晰布局、目标不密集场景	通过Census变换值表现整体结构

2.2 基于语义的分类方法

基于特征的方法虽然在一般场景分类任务中取得了良好的效果,但随着场景种类、复杂度的增加,其局限性也逐渐变得明显. 因此,对图像特征建模成为场景分类研究的新重点,下面就几种典型的语义分类方法进行分析与总结.

在 2001 年,Thomas Hofmann 等人认为如果若干词条多次出现在同一对象中,则这些词条在语义上具有相似性,随后提出了 LSA(Latent Semantic Analysis)[12]. 该方法首先应用在文本分析中,通过使用词与文档矩阵映射来描述词语是否处在文档中. 随后这种思想迅速沿用于场景分类技术当中,先对图像进行规则划分,提取各子块的局部图像描述子,建立局部描述子和语义概念之间的联系,再利用局部语义概念的概率分布完成基于图像场景语义的分类.

2010 年 Jia Li 等人认为场景是由一系列目标和结构构成的,有效识别目标语义成为进一步确定场景类别的前提,并通过收集大量的物体检测器在图像的不同空间位置上的多尺度响应来形成特征表示形成了 OB(Object Bank)方法[13]. OB 方法在室内复杂场景中表现出极佳的分类效果,然而在场景比较简单的情况下,其优势不再突出. 该方法过于依赖模型训练出的目标检测子,使得目标检测子的优劣及与场景的匹配程度直接决定这种方法的有效性.

在 2012 年,Fereshteh Sadeghi 等人提出了 LPR(Latent Pyramidal Regions)图像表示方法[14],该方法经由 LSVM(Latent Support Vector Machine)[15]训练出若干图像目标检测子以及一定数量的 SIFT 描述子,并在此基础上选取一到三层空间结构形成金字塔模型即 SPM(Spatial Pyramid Matching)空间结构模型[16],进行级联计算得到 LPR 特征向量. 该方法分离了图像区域检测子与分类器,从而能够进一步优化和设计分类器,有效地减小了整体空间结构的影响. 其引入局部空间结构来表现图像信息的思想较好地解决了场景变化所带来的问题. 然而,由于模型训练的局限性,在面对复杂多样的数据时并不具有通用性,这也是这类方法普遍存在的缺陷.

在 2013 年,Mayank Juneja 认为场景是由一般目标以及抽象目标组成的,是场景中具有显著分辨力的代表性部分,对这些目标进行检测及描述就能根据其内在语义推断出其场景类别,基于此提出了一种基于语义的 BOP(Bag of Parts)方法[17]. 该方法在对一幅图像分割和选择时,会过滤掉图像中包含的相似性信息,保留具有显著性差异的区域,不仅考虑并采集到了场景中常见的目标,同时也能将捕捉到抽象的目标特征. 基于这种思想的 BOW(Bag of Words)[18]等一系列模型方法都证明了一幅场景所包含的语义对于场景分类的重要性,目标选取更合理、模型更完善的 BOP 模型在图像场景分类中展现出更突出的效果,且由于其包含丰富多样的语义信息,成为图像场景分类中热门的模型之一.

综上所述,基于语义的图像场景分类方法着力于构建图像的语义层面的表示,实现了场景目标到语义概念的映射,逐渐成为当前图像场景分类的主流方法,具有较好的应用前景. 但由于其自身的条件限制,目前还存在诸多问题亟待优化. 语义方法的优缺点如表 2 所示.

表 2　图像语义方法比较

名称	优点		缺点	适用场合	特点
Object Bank[13]	可识别目标、自然场景、室内场景		计算复杂度高、特征维度高	标志性目标的自然场景,背景简洁的室内场景	通过多个目标检测子识别图像的标签
Latent Pyramidal Regions[14]	具有特定结构的区域表现好	适用背景复杂、目标拥挤的场景	侧重场景形状结构、缺乏深层语义理解	背景复杂、目标拥挤的室内场 景,依赖背景类别的抽象场景	通过区域检测子表现特定区域语义
Bag of Parts[17]	位置、边界、棱角表现好				通过超像素分割表现显著区域
Latent Semantic Analysis[12]	降维刻画同义词得到主题、冗余数据得到利用		多义词区分度低、计算复杂度高	信息驳杂,界限明显的室内场景	通过视觉词包确定对应空间信息

2.3 基于学习的分类方法

相较于传统的场景分类方法,利用卷积神经网络(CNN,Convolutional Neural Network)[19,20]进行场景分类具有三个优越性:首先,能够通过自学习的方式来“识别”图像;其次,利用反馈网络实现联想存储功能,可以更好地联想到图像所代表的具体含义;最后,可以发挥计算机的高速运算能力,更好地实现分类.

1986 年 David E Rumelhart 等人提出了多层神经网络权值修正的 BP(Back Propagation)算法[21],证明了其网络有很强的学习能力且能解决很多实际问题. 随着

GPU 和分布计算的发展使得计算能力大大提升,场景分类技术上也有了进一步的发展. 2006 年,Geoffrey Hinton 等人提出了深度学习的概念. 它的出现解放了人工提取特征的繁重劳役,对场景中的特征可自动进行提取及整合,在大型数据集下的运算和性能评估[22]有了很大的支持. 目前已有多种学习框架被提出及使用,如 Caffe[23]、Theano[24]、TensorFlow[25]、MXNet[26]、CNTK[27]等.

由于目标识别任务的复杂性,普通的机器学习方法不能够在诸如 ImageNet[28]这样的大型数据集上很好的运行,并且需要有先验知识来处理这些数据. 而 CNN 则可以通过改变深度和广度来控制自身的能力,对图像的性质有很好的评估准度. CNN 具有稀疏连接和权值共享的优点,减少了训练时网络参数的数目,降低了网络的复杂性,且多维信号的输入避免了特征提取和分类过程中的数据重排的过程,使得 CNN 在提高图像分类性能方面取得了巨大的成功[29,30].

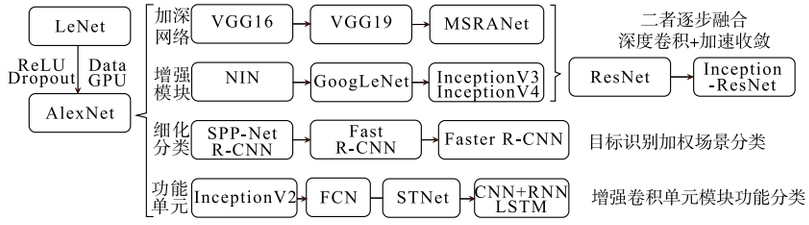

图1 场景分类中基于 CNN 模型的技术发展路线

如图 1 所示,经历了 LeNet 的启发及 GPU 和相关领域的兴起,2012 年,Alex Krizhevsky 等人创造了“大型的深度卷积神经网络”,即 AlexNet[31]. 随着技术的发展和数据的增长,基于场景分类的 CNN 模型逐渐分化为四个方向:加深网络结构、增强卷积模块、分类细化检测以及构架功能单元. 在 2014 年,VGGNet[32]的提出标志着加深网络结构模型的开始,其考量到 CNN 的深度与性能间的关系,构造了 16 ~ 19 层深的 CNN,相较于之前的网络结构,错误率大幅下降. 随后,部分网络结构采用增强其中的模块来提高 CNN 效率与准确度,具有代表性的 GoogLeNet[33]提出了 Inception 结构,其基本思想来源于 Shuicheng Cheng 的 NIN(Network in Network)[34],即原来的节点也是一个网络. InceptionV3 及衍生结构在单层卷积层上使用不同尺度的卷积,使得单层的特征提取能力变强,整个网络结构的宽度和深度扩大,并且减少了参数的数量,带来了 2 ~ 3 倍性能的提升. 目前,网络加深与模块增强呈现出融合的趋势,以达到深度卷积和加速收敛的目的,如 ResNet[35],它运用了残差网络结构,引入了批标准化和跨层连接,即通过在输出和输入之间引入一个 shortcut 链接,解决了网络由于深度过深而出现的梯度消失问题. 随着 Fast R-CNN[36]系列目标检测算法的兴起,场景分类领域随即引入并发展了细化分类的思想:通过以对象为单位,识别一些有代表性的对象来确定对象的位置和类别[37],继而更好地判断场景标签所属,达到目标识别加权场景分类的效果. 随后,自 Inception 模块的提出,CNN 的一部分焦点转移到了构架独立的功能单元,陆续提出了 FCN,ST-Net 等各具特点的结构,场景分类技术从多个角度进一步得到了优化和性能的提升.

深度学习网络在应用于场景分类问题时,一般情况下,其网络结构在逻辑上分为特征提取和分类两个阶段. 如图 2 所示,在特征提取部分,首先需要将经过处理的测试图片作为已训练好的 CNN 的输入,通过卷积层对图片的特征进行提取,再经过池化操作保留图片最显著的特征,并提升模型的畸变容忍能力. 每层的卷积与池化提取出的特征不同,如第一层提取出边缘信息,第二层提取出对象信息等. 但随着层数的增加,网络可以提取出更深层次的特征,而这些特征也具有了相应的语义信息. 之后进入分类阶段,将提

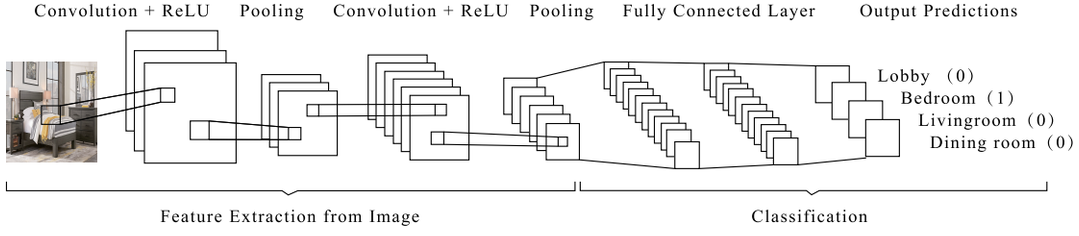

图2 场景分类中的 CNN 结构

取到的图像特征经过连接层的映射,得到一个向量,通过输出层 Softmax,判断其属于各类场景的概率并以此作为此图像的标签,并完成分类. 此外,深度学习网络也可以只作为特征提取的工具,抽取连接层的响应值作为图像的特征向量,再通过传统的分类器对其进行分类[38].

在不同方法的驱动下,一系列基于经典模型的改进方法不仅达到了有效的分类效果,也为后续研究人员的工作打开了新思路. Jing Sun 等人为了提高场景分类的精度,提出了一种新的场景分类算法[39],他们用 AlexNet 模型学习场景图片的特征,并提取 AlexNet 模型的最后一层作为图像特征,使用 Lib-SVM 训练出新的场景分类模型,使其具有更好的泛化能力,达到更好的分类精度. Cheng Gong 等人使用 AlexNet、VGGNet 和 GoogLeNet 三种 CNN 模型提取了高分辨率的遥感图像特征,使用 SVM(Support Vector Machine)分类器对图像进行了场景分类,并进行了特征有效性的对比[40]. Guoli Wang 等人提出了从 CNN 的不同层级信息中提取有识别性图像特征的新方法 EMR(Encoded Mixed-resolution Representation)[41]. 这种方法可以处理不同尺寸的图片,同时在 ImageNet 数据集上通过 VGGNet 和 ResNet 模型直接提取特征,可以减少在训练阶段的计算消耗,使其在场景分类上达到很好的效果. 在之后场景分类技术的发展过程中,研究者们通过不断的尝试,解决在其发展中所带来的问题.

3　数据库的描述及方法比较

3.1　数据库描述

一个优秀的数据集,在考虑规模、精度、多样性和层次结构等条件的同时,更要兼具完备的标准数据、设定参数和评价方式等特点. 近年来,根据研究焦点和工业生产需求,数据集呈现出多类别、多场景归一化的特性. 使得不同方法可以在同一数据集中通过统一评价方法进行比较.

随着场景分类发展,出现了许多数据集(如表 3 所示),详述如下.

COCO[44]:2016 年由 Tsung-Yi Lin 等人建立. 该数据集由微软赞助,其对于图像的标注信息极为丰富,不仅有类别、位置信息,还有对图像的语义文本描述. 该数据集的开源使得近几年图像分割,语义理解取得了巨大的进展,也几乎成为了图像语义理解算法性能评价的标准数据集.

SUN[45]:2015 年由 Jianxiong Xiao 等人提出,其中包含室外场景、自然场景、室内场景三个大类,是一个覆盖场景、位置、人物变化较大的数据库.

表 3　场景分类数据库(类别单位:类; 数量单位:万幅)

数据集	场景				描述		时间
	自然	城市	事件	室内	类别	数量	
WebVision[42]	√	√	√	√	约 1000	约 240	2017
Places[43]	√	√		√	约 400	约 1000	2017
COCO[44]	√	√	√		约 80	约 30	2016
SUN[45]	√	√		√	908	约 13	2015
ImageNet[28]	√	√	√	√	约 22000	约 1400	2012
MIT Indoor[48]				√	67	1.562	2009
CIFAR[49]	√	√		√	200	约 12	2009
UIUC Sports[50]	√	√	√		8	约 0.158	2007
LabelMe[51]	√	√		√	183	约 3	2008
LSP[16]	√	√		√	15	0.4485	2006
Pascal VOC[52]	√	√		√	20	1.1530	2005
LP[16]	√			√	13	0.3859	2005
OT[16]	√				8	0.2688	2001

ImageNet[28]:2012 年,由 Li Fei-Fei 及其团队构建. 其目的在于让机器学习避免过拟合并尽量满足更多实例,该数据集视觉信息复杂、模型趋于高维、并配以大量参数,是一个包含类别极多、内容丰富的大尺度数据集. 与之相关联的一年一度的比赛成为了每年场景分类研究者关注的焦点,赛程中往往包含着场景分类中最前沿的研究成果[46,47].

MIT Indoor[48]:由 2009 年 Li Fei-Fei 及其团队提出. 该数据集提供了包含丰富的局部及全局判别信息的图像,从而为当时应用于室内领域的模型方法提供了训练对象. 目前,该数据库已经成为室内图像场景分类的权威数据集之一.

Pascal VOC[52]:在 2005 年由 Mark Everingham 等人建立,作为一个供机器识别和训练的大型图片数据库,共包含 20 个大类别,每类图片数量在一千至一万张不等. 还衍生出针对于视觉对象的分类识别和检测的一个基准测试——PASCAL VOC 挑战赛,并提供了检测算法和学习性能的标准图像注释数据集和标准的评估系统.

除此之外,还有一些未被公开的数据集,包含了医疗、金融、生物等分支形成的图像信息,因其对专业知识极高的要求,相应的,获取成本和标注资源也在随之成倍递增. 研究组根据研究类别和需要,通过 Google 等搜索引擎临时建立的数据集,或通过监控器、无人机等通讯传输存储设备截取到的图片组建的数据集等,具有类别广泛、信息丰富、使用灵活、成本较低等特点. 如今,越来越多的研究者开始关注利用类似的低成本数据集(比如不含人工注释的数据)来训练图像识别系

统,并取得了良好的效果.

3.2 分类效果对比

本章节基于研究者高频使用的数据集和应用于室内、室外、自然和事件四个场景类别数据集的前沿方法进行了分析.我们对三类场景分类方法在不同场景的数据库上的分类效果进行对比和总结.每一类场景比较中,我们都列出了基于不同特征、语义及学习的不同分类方法在某几个特定场景数据库中的分类效果,并进行对比.

(1)基于特征的场景分类,如表4~7所示.

表4 基于特征的场景分类在室外场景的比较

场景类别	基于特征的分类方法	分类器	数据集			精度
			LSP	Google	UIUC Sports	
室外场景	GIST[14]	SVM			√	96.04%
	GIST[53]			√		94.22%
	CENTRIST[11]		√			73.30%
	SIFT、Pixel Grayvalue[54]	K-means	√			83.67%
	SIFT[16]		√			70.47%
	CENTRIST[55]	SPM	√			83.88%

表5 基于特征的场景分类在室内场景 MIT Indoor 数据集的比较

场景类别	基于特征的分类方法	分类器	精度
室内场景	HOG[56]	SVM	65.02%
	SIFT、HOG、GIST[55]		64.00%
	GIST[57]		61.30%
	CENTRIST[11]		36.90%
	SIFT[16]		33.33%
	ROI + GIST[3]		26.05%
	GIST[3]		22.00%

表6 基于特征的场景分类在事件场景数据集 UIUC Sports 的精度比较

场景类别	基于特征的分类方法	分类器	精度
事件场景	SIFT[16]	SVM + HIK	86.11%
	GIST[3]	SVM	63.90%

在室外场景中,我们在LSP、Google和UIUC Sports数据集上进行了对比,如表4所示.GIST特征的相关方法[14,53]中,文献[14]选取UIUC Sports数据集中部分图像测试GIST取得了96.04%良好效果.而有关于SIFT特征的方法[16,54],文献[54]结合灰度值及K-means分类器的模型有效地利用了各自的特点,在LSP数据集取得了83.67%的效果.CENTRIST特征与不同分类器结合呈现不同的效果[11,55],由于SPM金字塔型结构与CENTRIST特征对位置信息的良好表现力使其结合取得了更好的效果.

表7 基于特征的场景分类在自然场景的比较

场景类别	基于特征的分类方法	分类器	数据集		精度
			OT	Corel	
自然场景	局部 GIST[55]	SVM	√		86.85%
	SIFT[59]		√		86.22%
	全局 GIST[57]		√		83.55%
	SIFT[58]	SVM + HIK	√		87.80%
	RGB-SIFT[60]	K-means		√	82.28%

在室内场景中,我们在MIT Indoor数据集上进行对比,如表5所示.HOG特征的相关方法[55,56]都取得了较好的分类效果,基于HOG特征辅以SVM分类器能够对场景的边缘形状和目标轮廓进行描述,在MIT Indoor主流室内数据集上达到相对较好的结果65.02%[56].由于不同特征本身的局限性,对于开放程度较小、局部信息丰富的室内场景,基于CENTRIST特征方法[11]效果不佳,基于GIST[3]、SIFT[16]特征方法依赖于全局结构使得其分类效果较差.

在事件场景中,我们在UIUC Sports数据集上进行对比(表6).由于基于特征的方法对于事件场景表现力欠佳,基于SIFT的分类方法[16]通过结合SVM、HIK模型的同时融入了标签信息,保证了在子空间建模被描述的均衡性,使得效果达到86.11%,基于GIST的方法[3]由于事件场景通常信息复杂多元,分类效果仅为63.90%.

在自然场景中,我们在OT、Corel数据集上进行对比,如表7所示.颜色信息对于图像场景分类十分重要,文献[57]忽略了场景的重要特性,分类效果仅达到83.55%,文献[55]引入颜色信息并考虑到了图像的局部特征,结合了其固有的全局特性,使得其在SVM分类器下达到了较好的效果86.85%.SIFT特征作为一种基于梯度的局部特征图像块描述子,其分类方法[58]普遍适用于自然场景,结合了SVM、K-means分类器,分类效果相差不多.通过SVM与HIK分类器之间特征的抽象化联合表示以及SIFT描述子与自然场景特点的契合度,分类效果较好达到87.8%.

(2)基于语义的场景分类,如表8~11所示.

在自然场景中,我们在OT、Corel、LabelMe数据集上进行了对比(表8).Jianzhao Qin等人提出基于多尺度稀疏表示的场景分类框架[61],并结合不同尺度上的语义信息进行串接,汇总为图像的全局向量,在OT库达到了88.81%的良好效果.由于多层次描述复杂度较高,出现了一系列基于局部语义信息处理的方法[62,63],分类精度相差不多.而基于MIML的多标签学习方法[64,65]依赖于数据集本身的标注情况和图像质量,虽结合不同的分类器但分类效果仍不佳.

在室内场景中,我们在 MIT Indoor 数据集上进行了对比,如表 9 所示. 由于场景丰富多变、结构复杂的特点,文献[17]的分类效果虽然仅达到了 60.80%,但其强调了语义信息的提取和分析的观点,提出的一系列评价函数(entropy-rank)和模型方法(BOW、Fisher vector model),为后续的模型奠定了坚实的基础,引领了当时的研究热点. 此后 Parizi 等人提出的 RBOW 方法[66]能够重组若干区域进行集合表示,Zhang 等人提出的 OB 方法[13]能够集中表现场景中有表现力的区域,但都在室内场景中分类效果不佳. 对于更加注重模型与目标匹配的 DPM 方法[10],对室内场景的形变问题表现不佳,分类效果较差.

表 8　基于语义的场景分类在自然场景的比较

场景类别	基于语义的分类方法	分类器	数据集			精度
			OT	Corel	LabelMe	
自然场景	MIML[64]	SVM			√	77.67%
	ML[62]	I2C			√	82.08%
	SIFT、NCH、BOW[60]	K-means		√		82.28%
	MIML[65]	Boost			√	78.33%
	Qin[61]		√			88.81%
	InsDif[63]				√	82.98%

表 9　基于语义的场景分类在室内场景 MIT Indoor 数据集的比较

场景类别	基于语义的分类方法	分类器	精度
室内场景	IFV[17]	SVM	60.80%
	LLC[17]		53.00%
	RBOW[66]		37.90%
	BOP[17]		46.10%
	OB[13]		37.60%
	DPM[10]		30.40%

在室外场景中,我们在 LSP、LP、UIUC Sports、Google 数据集上进行对比,如表 10 所示. 针对于图像模糊化的全局语义的基于 LBP 与 SVM 的 ST 算法[53]被提出,它无需目标局部信息,能够用较少的维数包含多种语义信息,同时可以兼顾多层上下文图像信息,虽然在 UIUC Sports 数据集上分类效果达到 96.65%,且在 Google 数据集上同样表现出了不错的效果,但受限于场景要素的数量,缺少对不同图像的泛化能力. 基于 BOF 的方法[67]没有使用主流分类器,而是利用整体结构特征和局部纹理特征,结合两级分类器在室外场景中进行分类,达到了很好的效果 86.06%. 而基于局部语义以及特征结合的分类方法[67,68]由于关键性区域在室外场景中无法完全体现出其特性,分类效果表现不佳. 对于一系列基于 SPM 的改进方法[16,67,69],通过将语义信息进一步提取并组合,分类效果相较区域性方法有所提升.

表 10　基于语义的场景分类在室外场景的比较

场景类别	基于语义的分类方法	分类器	数据集				精度
			LSP	LP	UIUC Sports	Google	
室外场景	LBP[53]	SVM			√		96.65%
						√	94.55%
	SIFT、BOW、PLSA[68]			√			78.10%
	BOF[67]		√				86.06%
	sPACT[67]		√				83.30%
	ScSPM + HIK[69]		√				82.40%
	KSPM[16]		√				81.40%
	SPM[16]		√				81.40%
	ScSPM[67]		√				80.28%
	KC[67]		√				76.67%
	Dai[58]		√				75.70%

在事件场景中,我们在 UIUC Sports 数据集上进行了对比,如表 11 所示. 基于 LPR 的分类方法[14]通过引入用户标记过程,融合互补多组特征形成多视点的降维模型,取得了 86.50% 的良好分类效果. 而基于 SPM 的改进模型[16,69]没有考虑低维空间构造中不同视角的多个特征.

表 11　基于语义的场景分类在事件场景数据集 UIUC Sports 的比较

场景类别	基于语义的分类方法	分类器	精度
事件场景	SIFT、NCH、BOW[60]	K-means	83.98%
	LPR[14]		86.50%
	Qin[61]		85.90%
	ScSPM[69]		82.74%
	OB[13]		76.30%
	SPM[16]		71.60%

(3)基于学习的场景分类,由于数据量大、数据集的不断迭代更新,使得数据集中一些场景之间的区分性并不明显,比如一些图像可能同时包含了室内场景和事件场景的信息等. 如在表 12 所示的数据集中,ImageNet 和 Places 作为场景分类中的大型数据库包含了众多不同类型、细致化的场景,AlexNet 与 VGGNet-16 方法分别在这两个数据集上进行了对比[70]. VGGNet-16 网络相对于 AlexNet 网络,使用 3×3 的卷积核和 2×2 的池化核,通过不断加深网络结构提升了性能,在场景分类中达到更好的效果,在 ImageNet 和 Places 数据集准确率分别达到 73.00% 和 60.60%. 在文献[36]中,Places-CNN 特征、ImageNet-CNN 特征分别与 SVM 分类器相结合,并在 SUN、MIT Indoor 和 LSP 数据库上进行

了测试,由表中数据可知,由于 LSP 数据库数据量相对较少,所以分类精准度较高. 而 Hybrid-CNN 网络结合 Places-CNN 特征、ImageNet-CNN 特征进一步提取新的特征,因而在 LSP 数据库中的分类效果最佳. Yangzihao Wang 等人[71]提出的网络结构,通过预测建议区域,结合 CNN 网络的特点进行场景分类,与 Places-CNN 和 ImageNet-CNN 特征方法在 MIT Indoor 数据集上进行了对比,效果较好 68.30%. 总之,在基于学习的场景分类方法中,所追求的目标已转变为方法的泛化性.

表 12 基于学习的场景分类的比较

基于学习的分类方法	数据集				
	ImageNet	Places	SUN	MIT Indoor	LSP
AlexNet[70]	59.30%	50.00%			
VGGNet-16[70]	73.00%	60.60%			
Places-CNN[36]			54.32%	68.24%	90.19%
ImageNet-CNN[36]			42.61%	56.79%	84.24%
Hybrid-CNN[36]			53.86%	70.80%	91.59%
Yangzihao Wang[71]				68.30%	

4 未来的挑战与趋势

随着海量数据下对场景自动分类需求的不断增长,场景分类技术面临着不断的挑战,所面临问题及发展趋势如下:

4.1 场景分类面临的挑战

(1)大规模复杂场景分类:随着图像数据的海量增长及类别的不断细分,使得场景分类所面临的问题无论是在图像数量上还是在图像类别上都面临着前所未有的挑战.

(2)视频场景分类:视频数据比单帧图像包含更多元更有效的信息,即使视频中某一场景特征不变其他特征也会在不断地发生改变,因此视频的出现,既是机遇也是挑战.

(3)真实场景理解:在实际应用中,场景的理解不再是单纯的分类问题,更多的是对内容语义的理解和实时性要求. 真实场景的理解在复杂度和深度上都比场景分类要高,此问题解决后将对人工智能技术产生深远的影响.

4.2 场景分类的发展趋势

场景分类所带来的挑战,势必会继续带动算法和技术的发展,为新思路的出现做铺垫. 就目前所面临的问题来看,场景分类的发展趋势主要有以下几个方面.

(1)场景分类技术的发展

基于多源特征提取:目前的特征提取方法大多是通过目标检测实现的,忽略了场景内其他信息的作用,这尤其对小数据样本以及特殊场景分类影响较大. 因此,综合目标和背景等进行多源特征的提取将成为下一步发展以及研究趋势.

基于深度学习的泛化分类:随着场景数据集趋向复杂化、分类趋向多元化,深度学习对于大数据具有良好的分析与表现效果,尤其是面对目前呈现爆炸式增长的多媒体数据,能够根据需求自动地优化分类决策和类别数量. 在当今的大数据时代,基于深度学习的分类方法已经普遍应用于场景分类当中,但只能通过特殊的训练集来区分场景. 基于深度学习泛化研究则是通用场景分类研究的趋势和焦点.

(2)场景分类应用的发展

场景分类是对场景图像进行粗分类,而基础的分类已经不能够满足当今社会对于场景的理解. 场景理解作为场景分类的进一步发展和应用,对场景中各个目标与对象进行理解,对多个模型进行信息融合,对图像中的特征进行深度分 析,实现类似人类视觉系统理解外部世界的过程. 场景分类作为基础,将在场景理解中起着至关重要的作用.

5 总结

本文针对场景分类对国内外场景分类发展的现状及前沿成果进行了论述. 对基于不同理念和技术的场景分类方法加以分类和深入挖掘,对于各类别所属的前沿方法进行系统的归纳、分析和总结,最后对场景分类技术存在的问题和未来发展趋势做出了展望. 如今,场景结构复杂多变,场景内容千差万别,如何利用计算机来模拟人脑的思维方式成为关键,而这恰恰也是计算机视觉研究中最困难的地方. 在层出不穷的挑战与难题的衍生与攻克的过程当中,场景分类将会在更广泛的环境和数据下趋向于自适应性的分类,同时也为场景理解的研究奠定了基础. 因此,在未来的场景分类问题中,需要更多来自于多学科的学者进行交叉讨论与研究,以取得进一步的突破.

参考文献

[1] ZHEN L, LIU Q. A new feature selection method for internet traffic classification using ml[J]. Physics Procedia, 2012, 33(2): 1338 – 1345.

[2] 李学龙,史建华,董永生,陶大程. 场景图像分类技术综述[J]. 中国科学:信息科学,2015,45(7):827 – 848.
LI Xue-long, SHI Jian-hua, DONG Yong-sheng, TAO Da-cheng. A survey on scene image classification[J]. Science China: Information, 2015, 45(7): 827 – 848. (in Chinese)

[3] OLIVA A, TORRALBA A. Modeling the shape of the scene: A holistic representation of the spatial envelope[J]. International Journal of Computer Vision, 2011: 145 – 175.

[4] FELZENSZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA: IEEE, 2008. 1 – 8.

[5] LI J, WANG J-Z, GRAY R-M, WIEDERHOLD G. Multiresolution object-of-interest detection for images with low depth of field[A]. ICIAP[C]. Washington DC, USA: IEEE Computer Society, 1999. 32 – 37.

[6] SHEN J, SHEPHERD J, NGU A-H-H. Semantic-sensitive classification for large image libraries[A]. Proceedings of International Multimedia Modelling Conference[C]. Piscataway, USA: IEEE, 2005. 340 – 345.

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

[8] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA: IEEE, 2005. 886 – 893.

[9] BOSCH A, ZISSERMAN A, MUNOZ X. Representing shape with a spatial pyramid kernel[A]. ACM International Conference on Image and Video Retrieval[C]. New York, USA: Association for Computing Machinery, 2007. 401 – 408.

[10] PANDEY M, LAZEBNIK S. Scene recognition and weakly supervised object localization with deformable part-based models[A]. Proceedings of IEEE International Conference on Computer Vision[C]. Piscataway, USA: IEEE, 2011. 1307 – 1314.

[11] WU J, REHG J-M. CENTRIST: A visual descriptor for scene categorization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8): 1489 – 1501.

[12] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning, 2001, 42(1 – 2): 177 – 196.

[13] LI L-J, SU H, XING E-P, FEI-FEI L. Object bank: A high-level image representation for scene classification & semantic feature sparsification[A]. Proceedings of Conference on Neural Information Processing Systems[C]. Canada: Neural Information Processing System Foundation, 2010. 1378 – 1386.

[14] SADEGHI F, TAPPEN M-F. Latent pyramidal regions for recognizing scenes[A]. Proceedings of European Conference on Computer Vision[C]. Germany: Springer International Publishing, 2012. 228 – 241.

[15] ANDREWS S, TSOCHANTARIDIS I, HOFMANN T. Support vector machines for multiple-instance learning[A]. Proceedings of Conference on Neural Information Processing Systems[C]. Canada: Neural Information Processing System Foundation, 2003. 561 – 568.

[16] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[A]. Proceedingsof Computer Vision and Pattern Recognition[C]. Piscataway, USA: IEEE, 2006. 2169 – 2178.

[17] JUNEJA M, VEDALDI A, JAWAHAR C, ISSERMAN A. Blocks that shout: Distinctive parts for scene classification[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA: IEEE, 2013. 923 – 930.

[18] WALLRAVEN C, CAPUTO B, GRAF A. Recognition with local features: the kernel recipe[A]. Proceedings of IEEE International Conference on Computer Vision[C]. Piscataway, USA: IEEE, 2003. 1 – 8.

[19] RAZAVIAN A-S, AZIZPOUR H, SULLIVAN J, CARLSSON S. CNN features off-the-shelf: An astounding baseline for recognition[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA: IEEE, 2014. 512 – 519.

[20] HERSHEY S, CHAUDHURI S, ELLIS D-P, et al. CNN architectures for large-scale audio classification[A]. ICASSP[C]. Piscataway, USA: IEEE, 2017. 131 – 135.

[21] CRICK F, ASANUMA C, MCCLELLAND J, RUMELHART D. Parallel distributed processing: Explorations in the microstructure of cognition[J]. Psychological and Biological Models, 1986, 63(4): 45 – 76.

[22] HINTON G-E, OSINDERO S, THE Y-W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527 – 1554.

[23] JIA Y, SHELHAMER E, DONAHUE J, et al. CAFFE: Convolutional architecture for fast feature embedding[A]. ACM International Conference on Multimedia Retrieval[C]. Piscataway, USA: IEEE, 2014. 675 – 678.

[24] BERGSTRA J, BREULEUX O, BASTIEN F, et al. Theano: a CPU and GPU math expression compiler[A]. Proceedings of Python in Science Conference[C]. Piscataway, USA: IEEE, 2010. 3 – 10.

[25] ABADI M, AGARWAL A, BARHAM P, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems[A]. ARXiv Preprint[C]. 2016. arXiv: 1603. 04467.

[26] CHEN T, LI M, LI Y, et al. MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems[A]. Proceedings of Conference on Neural Information Processing Systems[C]. Canada: Neural Information Processing System Foundation, 2015. 1 – 6.

[27] YU D, EVERSOLE A, SELTZER M, et al. An Introduc-

tion to Computational Networks and the Computational Network Toolkit[M]. USA:Microsoft Research,2014.
[28] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA:IEEE,2009. 248 – 255.
[29] MARGOLIN R, ZELNIK-MANOR L, TAL A. OTC: A novel local descriptor for scene classification[A]. Proceedings of European Conference on Computer Vision[C]. Germany: Springer International Publishing, 2014. 377 – 391.
[30] OQUAB M, BOTTOU L, LAPTEV I, SIVIC J. Learning and transferring mid-level image representations using convolutional neural networks[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA:IEEE,2014. 1717 – 1724.
[31] KRIZHEVSKY A, SUTSKEVER I, HINTON G-E. ImageNet classification with deep convolutional neural networks[A]. Proceedings of Conference on Neural Information Processing Systems[C]. Canada: Neural Information Processing System Foundation,2012. 1097 – 1105.
[32] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[A]. ArXiv preprint[C]. 2014. arXiv:1409. 1556.
[33] SZEGEDY D-C, LIU W, JIA Y, et al. Going deeper with convolutions[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA: IEEE, 2015. 1 – 9.
[34] LIN M, CHEN Q, YAN S. Network in network[A]. arXiv preprint[C]. arXiv:1312. 4400,2013.
[35] HE K, ZHANG X, REN S, SUN J. Deep residual learning for image recognition[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA: IEEE,2016. 770 – 778.
[36] GIRSHICK R. Fast R-CNN[A]. Proceedings of IEEE International Conference on Computer Vision[C]. Piscataway, USA:IEEE,2015. 1440 – 1448.
[37] TIAN Y, ZHANG W, ZHANG Q, LU G. Inception classification and object detection based joint-CNN for indoor scene classification[A]. Proceedings of International Multi Conference on Engineers and Computer Scientists[C]. Germany: Springer International Publishing, 2018. 334 – 338.
[38] DONAHUE J, JIA Y, VINYALS O, et al. Decaf: A deep convolutional activation feature for generic visual recognition[A]. Proceedings of International Conference on Machine Learning[C]. Madison, USA: Omnipress, 2014. 647 – 655.
[39] SUN J, CAI X, SUN F, ZHANG J. Scene image classification method based on AlexNet model[A]. Proceedings of Informative and Cybernetics for Computational Social Systems[C]. Piscataway, USA:IEEE,2016. 363 – 367.
[40] CHENG G, MA C, ZHOU P, et al. Scene classification of high resolution remote sensing images using convolutional neural networks[A]. Proceedings of IEEE International Symposium Geoscience and Remote Sensing[C]. Piscataway, USA:IEEE,2016. 767 – 770.
[41] WANG G, FAN B, XIANG S, PAN C. Aggregating rich hierarchical features for scene classification in remote sensing imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2017,10(9):4104 – 4115.
[42] LI W, WANG L, LI W, et al. Webvision database: Visual learning and understanding from web data[A]. ARXiv Preprint[C]. 2017. arXiv:1708. 02862.
[43] ZHOU B, LAPEDRIZA A, KHOSLA A, et al. Places: a 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,40(6):1452 – 1464.
[44] LIN T-Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[A]. Proceedings of European Conference on Computer Vision[C]. Germany: Springer International Publishing,2014. 740 – 755.
[45] XIAO J, HAYS J, EHINGER K-A, et al. Sun database: Large-scale scene recognition from abbey to zoo[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA:IEEE,2010. 3485 – 3492.
[46] RUSSAKOVSKY O, DENG J, HUANG Z, et al. Detecting avocados to zucchinis: What have we done, and where are we going? [A]. Proceedings of IEEE International Conference on Computer Vision[C]. Piscataway, USA:IEEE,2013. 2064 – 2071.
[47] RASTEGARI M, ORDONEZ V, REDMON J, FARHADI A. XnorNet: ImageNet classification using binary convolutional neural networks[A]. Proceedings of European Conference on Computer Vision[C]. Germany: Springer International Publishing,2016. 525 – 542.
[48] QUATTONI A, TORRALBA A. Recognizing indoor scenes[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA: IEEE, 2009. 413 – 420.
[49] KRIZHEVSKY A. Learning Multiple Layers of Features From Tiny Images[R]. Technical Report,2009,1(4):1 – 58.
[50] LI L-J, LI F-F. What, where and who? classifying events by scene and object recognition[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA:IEEE,2007. 1 – 8.

[51] RUSSELL B-C, TORRALBA A, MURPHY K-P, FREEMAN W-T. LabelMe: A database and web-based tool for image annotation[J]. International Journal of Computer Vision, 2008, 77(1-3): 157-173.

[52] EVERINGHAM M, GOOL L-V, WILLIAMS C, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[53] 李锦锋,许勇. 基于 LBP 和小波纹理特征的室内室外场景分类算法[J]. 中国图象图形学报,2010,15(5):742-748.
LI Jin-feng, XU Yong. Indoor outdoor scene classification algorithm based on the texture feature of LBP and Wavelet[J]. Journal of Image and Graphics, 2010, 15(5): 742-748. (in Chinese)

[54] NAVARRO F, ESCUDERO-VIÑOLO M, BESCOS J. SP-SIFT: enhancing SIFT discrimination via super-pixel-based foreground-background segregation[J]. Electronics Letters, 2014, 50(4): 272-274.

[55] GEMERT J-C, GEUSEBROEK J-M, VEENMAN C-J, SMEULDERS W. Kernel codebooks for scene categorization[A]. Proceedings of European Conference on Computer Vision[C]. Germany: Springer International Publishing, 2008. 696-709.

[56] ZHANG Q, YANG J, ZHANG S. Indoor scene classification based on mid-level features[A]. Information Technology and Intelligent Transportation Systems[C]. Germany: Springer International Publishing, 2017. 235-242.

[57] OLIVA A, TORRALBA A. Building the gist of a scene: The role of global image features in recognition[J]. Progress in Brain Research, 2006, 155(2): 23-36.

[58] DAI D, YANG W, WU T. Three-layer spatial sparse coding for image classification[A]. Proceedings of International Conference on Pattern Recognition[C]. Germany: Springer International Publishing, 2010. 613-616.

[59] BOSCH A, ZISSERMAN A, MUÑOZ X. Scene classification via plsa[A]. Proceedings of European Conference on Computer Vision[C]. Germany: Springer International Publishing, 2006. 517-530.

[60] 崔鉴,段菲,章毓晋. 利用编码层特征组合进行场景分类[J]. 吉林大学学报:工学版,2013,43:450-454.
CUI Jin, DUAN Fei, ZHANG Yu-jin. Scene classification based on coding layer feature combination[J]. Journal of Jilin University (Engineering and Technology Edition), 2013, 43: 450-454.

[61] QIN J, YUNG N-H-C. Scene categorization with multi-scale category-specific visual words[J]. Optical Engineering, 2009, 48(4): 047203.

[62] WANG Z, HU Y, CHIA L-T. Multi-label learning by image-to-class distance for scene classification and image annotation[A]. Proceedings of the ACM International Conference on Image and Video Retrieval[C]. New York, USA: Association for Computing Machinery, 2010. 105-110.

[63] ZHANG M-L, ZHOU Z-H. Multi-label learning by instance differentiation[A]. Proceedings of National Conference on Artificial Intelligence[C]. Menlo Park, USA: American Association for Artificial Intelligence, 2007. 669-674.

[64] ZHOU Z-H, ZHANG M-L. Multi-instance multi-label learning with application to scene classification[A]. Proceedings of Conference on Neural Information Processing Systems[C]. New York, USA: Curran Associates, 2007. 1609-1616.

[65] ZHOU Z-H, ZHANG M-L, HUANG S-J, LI Y-L. Multi-instance multi-label learning[J]. Artificial Intelligence, 2012, 176(1): 2291-2320.

[66] PARIZI S-N. Reconfigurable models for scene recognition[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA: IEEE, 2012. 2775-2782.

[67] 程刚,王春恒. 基于结构和纹理特征融合的场景图像分类[J]. 计算机工程,2011,37(5):227-229.
CHENG Gang, WANG Chun-heng. Scene image categorization based on structure and texture feature fusion[J]. Computer Engineering, 2011, 37(5): 227-229.

[68] ZENG P, WU L, WEN J. Scene classification based on block latent semantic[J]. Journal of Computer Applications, 2008, 28(6): 1537-1542.

[69] YANG J, YU K, GONG Y, HUANG T. Linear spatial pyramid matching using sparse coding for image classification[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Piscataway, USA: IEEE, 2009. 1741-1801.

[70] WANG L, GUO S, HUANG W, et al. Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs[J]. IEEE Transactions on Image Processing, 2016, 26(4): 2055-2068.

[71] WANG Y, WU Y. Scene classification with deep convolutional neural networks[OL]. https://pdfs.semanticscholar.org/05b3/d0ece3c05be45b1ae7db3b43123befae3474.pdf, 2015.

作者简介

田艳玲 女,1992 年 10 月出生,山西省临汾人,现为陕西师范大学计算机科学学院硕士生. 主要从事机器视觉、场景分类等方面的研究工作.
E-mail:yl. tian@ siat. ac. cn

张维桐 男,1997 年 1 月出生,吉林省长春人,现为陕西师范大学计算机科学学院本科. 软件工程专业.
E-mail:weitong. zhang@ foxmail. com

张锲石(通讯作者) 男,1981 年 12 月出生,甘肃省兰州人,工学博士. 2014 年于日本早稻田大学获工学博士学位,现为中国科学院深圳先进技术研究院高级工程师,主要从事机器视觉、场景分类等方面的研究工作.
E-mail:qs. zhang@ siat. ac. cn

路 纲(通讯作者) 男,1972 年出生,四川省成都人,工学博士. 2009 年于电子科技大学获工学博士学位,现为陕西师范大学计算机科学学院副教授,主要从事无线自组织网络,智能信息融合,模式识别等方面的研究工作.
E-mail:goforlg@ 126. com

吴晓军 男,1970 年出生,陕西省宝鸡人,工学博士. 2005 年于西北工业大学获博士学位,现为陕西师范大学计算机科学学院教授.
E-mail:xjwu@ snnu. edu. cn