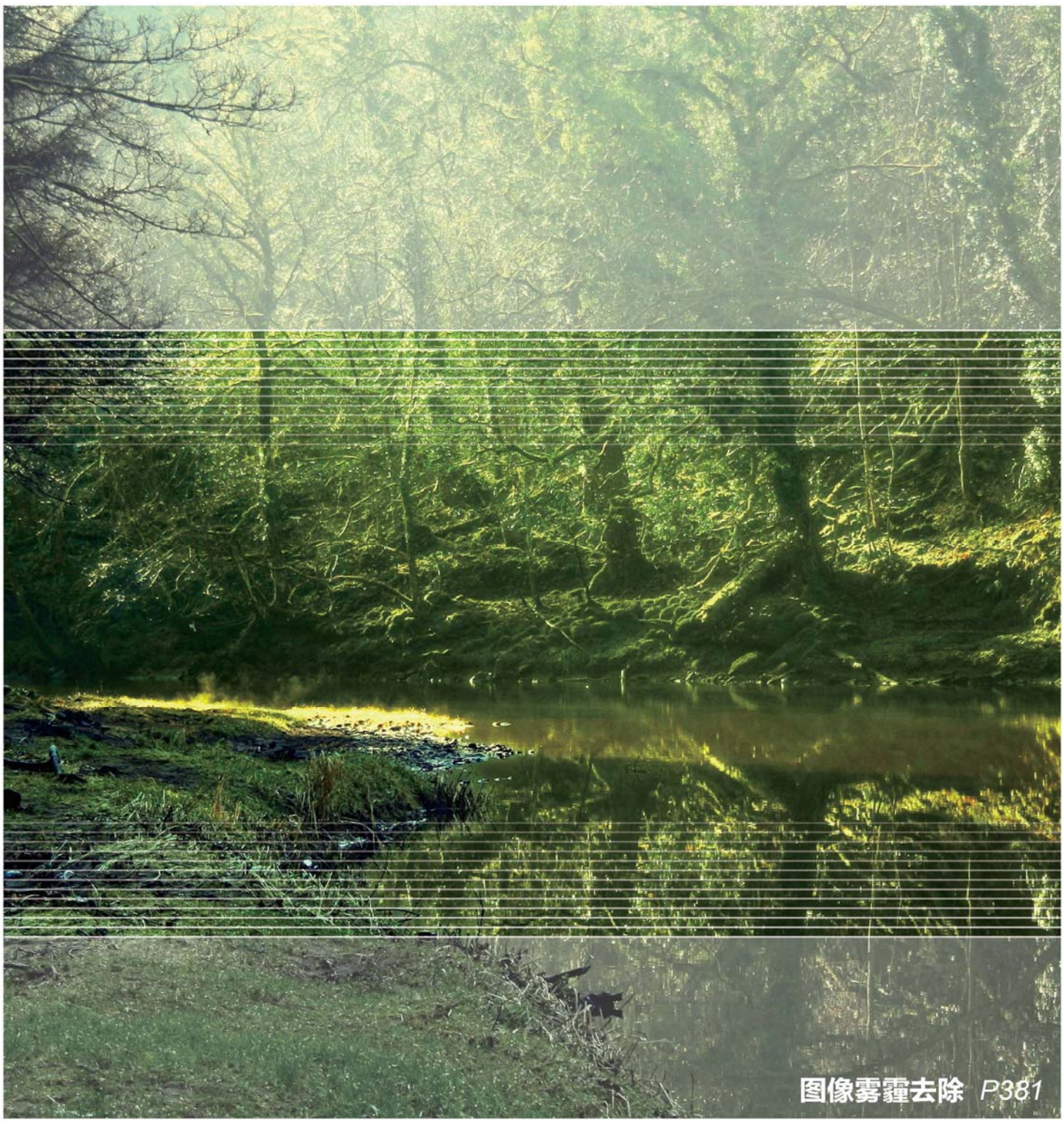
JIG
JOURNAL OF IMAGE AND GRAPHICS

中国图象图形学报

主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

2014
03
VOL.19

ISSN1006-8961
CN11-3758/TB



图像雾霾去除 P381

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第19卷第3期（总第215期）
2014年3月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 李小文
编辑出版 《中国图象图形学报》编辑出版委员会
邮政信箱 北京9718信箱
邮　　编 100101
电子信箱 jig@irsa.ac.cn
电　　话 010-64807995
网　　址 www.cjig.cn

广告经营许可证 京朝工商广字第0361号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief LI Xiaowen
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
P.O.Box 9718, Beijing, P.R.China
Zip code 100101
E-mail jig@irsa.ac.cn
Telephone 010-64807995
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 50.00元

2014年3月 第19卷 第3期（总第215期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

控制关键帧选择的H.264熵编码加密算法(第358页)

显著性检测指导的高光区域修复(第393页)

由灭点进行径向畸变的自动校正(第407页)

Volume 19, Number 3
Published March 16, 2014

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

H.264 video entropy coding encryption by controlling key frames(P358)

Algorithm for highlight area inpainting guided by saliency detection(P393)

Automatic approach of lens radial distortion correction based on vanishing points(P407)

E-mail: jig@irsa.ac.cn
Website: www.cjig.cn
Tel: 010-64807995

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号: TP391.41　　**文献标识码**: A　　**文章编号**: 1006-8961(2014)03-0333-11

论文引用格式: 赵理君,唐娉,霍连志,郑柯. 图像场景分类中视觉词包模型方法综述[J]. 中国图象图形学报,2014,19(3):333-343. [DOI: 10.11834/jig.20140301]

图像场景分类中视觉词包模型方法综述

赵理君[1,2], 唐娉[1], 霍连志[1], 郑柯[1]

1. 中国科学院遥感与数字地球研究所, 北京　100101; 2. 中国科学院大学, 北京　100049

摘　要: **目的**　关于图像场景分类中视觉词包模型方法的综述性文章在国内外杂志上还少有报导,为了使国内外同行对图像场景分类中的视觉词包模型方法有一个较为全面的了解,对这些研究工作进行了系统总结。**方法**　在参考国内外大量文献的基础上,对现有图像场景分类(主要指针对单一图像场景的分类)中出现的各种视觉词包模型方法从低层特征的选择与局部图像块特征的生成、视觉词典的构建、视觉词包特征的直方图表示、视觉单词优化等多方面加以总结和比较。**结果**　回顾了视觉词包模型的发展历程,对目前存在的多种视觉词包模型进行了归纳,比较常见方法各自的优缺点,总结了视觉词包模型性能评价方法,并对目前常用的标准场景库进行汇总,同时给出了各自所达到的最高精度。**结论**　图像场景分类中视觉词包模型方法的研究作为计算机视觉领域方兴未艾的热点研究领域,在国内外研究中取得了不少进展,在计算机视觉领域的研究也不再局限于直接应用模型描述图像内容,而是更多地考虑图像与文本的差异。虽然视觉词包模型在图像场景分类的应用中还存在很多亟需解决的问题,但是这丝毫不能掩盖其研究的重要意义。

关键词: 场景分类;视觉词包;低层特征;直方图表示

Review of the bag-of-visual-words models in image scene classification

Zhao Lijun[1,2], Tang Ping[1], Huo Lianzhi[1], Zheng Ke[1]

1. *Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing* 100101, *China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*

Abstract: **Objective** With the rapid development of computer multi-media technique, database technique and computer network technique, there have been more and more images to classify and label. Instead of using traditional manual mode, it has been a hot research field to use computer-aided automatic image-scene classification techniques. Among numerous image scene classification methods, the bag-of-visual-words (BOVW) model has become a widely adopted one, which, as a middle level feature, can narrow the gap between low-level visual features and high-level semantic features. However, reviews about BOVW model in image scene classification are rarely seen on journals in China and abroad. Therefore, in order to give a comprehensive understanding of this method to researchers in this field, this paper systematically summarizes these studies. **Method** Based on numerous references about the BOVW model in image scene classification during almost the past ten years, we divide the general process of development of the BOVW into five stages, that is, the stage of direct application of early bag-of-words model in image field, the stage of studying latent semantic information in the BOVW model, the stage of studying spatial layout or structure information in the BOVW model, the stage of studying context informa-

收稿日期:2013-07-11;**修回日期**:2013-09-10

基金项目:国家高技术研究发展计划(863)基金项目(2012AA12A304);中国科学院遥感与数字地球研究所所长青年基金项目(Y3SJ7700CX)

第一作者简介:赵理君(1986—),男,现为中国科学院遥感与数字地球研究所2012级博士研究生,主要研究方向为高分辨率遥感图像场景理解。E-mail:zhaolj01@radi.ac.cn

通信作者:唐娉,E-mail:tangping@radi.ac.cn

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 3, Mar. 2014

tion in the BOVW model, and the stage of optimizing visual word semantics and introducing new methods into the BOVW model. Furthermore, we sum up and compare various existing BOVW models in image scene classification in terms of local feature selection, feature generation of local image patches, visual vocabulary construction, histogram representation of bag of visual words feature, optimization of visual words, and so on. **Result** The development history of the BOVW and the research status of the BOVW based image scene classification are reviewed, which gives a clear trail of the development of the BOVW model; the numerous existing the BOVW models are categorized according to their working mechanism; the advantages and disadvantages of commonly used methods are compared; the performance evaluation method for the BOVW model is described and the commonly used standard scene databases are collected, with their best classification accuracies given separately. **Conclusion** As a hot research field that is currently rising, studies of the BOVW methods in image scene classification have produced quite a few research progress. The research in computer vision field has no longer been limited to directly applying original the BOVW model to describe image content, and more and more differences between images and texts are considered. The urgent problems to be solved are as follows: the performance of the BOVW will be greatly influenced when the bag of visual words are applied to the samples that are quite different from the training ones, while training new bag of visual words based on new training samples is very time and labor consuming; there is still no theoretical guide for determining the size of visual vocabulary; the relationship between visual words and semantics is still not fully exploited; the application of the BOVW in special fields, such as high-resolution remote sensing land-use scene classification, is far from being satisfactory. Besides, based on these problems, there may be some interesting research directions for example: constructing universal self-adaptive bag of visual words for different sample sets, automatically selecting optimal vocabulary size, adding more spatial layout and context information to the BOVW and exploring latent semantic information in visual words, studying image visual grammars for image understanding, studying scene classification problems in images of special fields, such as high resolution remote sensing images, and investigating new well-characterized low level feature extraction algorithms to construct high level bag of visual words. To conclude, although there are still a number of urgent problems to be solved in the application of the BOVW model based image scene classification, the important meanings of the studies of the BOVW model cannot be covered up.

Key words: scene classification; bag-of-visual-words; low-level feature; histogram representation

0 引 言

随着计算机多媒体技术、数据库技术和计算机网络技术的飞速发展,图像信息量迅猛增加,依靠人工对海量图像进行分类和标注的管理方式已经远远无法满足应用的需求。因此,利用计算机技术自动进行图像理解(image understanding)成为目前的一个研究热点,它主要是对图像的语义解释,研究图像中包含什么目标、目标之间的相互关系以及图像场景分类等问题[1]。其中,图像场景分类主要侧重于对图像场景整体的认识和分析,获得其自身的场景语义信息以解释图像场景的内容,而较少涉及包含的具体目标。为了得到对场景的语义描述,需要在监督训练的框架下建立高层场景语义描述与低层视觉特征之间的联系。早期关于场景分类方面的研究,通常采用颜色、纹理、形状等低层特征建立图像场景模型,利用分类器对场景的高层信息进行推导。然而,采用基于低层特征描述的场景分类方法由于缺乏中间语义的图像表示,所以泛化性差,很难用于处理训练集以外的图像。

为了克服图像低层视觉特征与高层语义之间的鸿沟,基于中层特征对场景语义建模描述的方法逐渐得到广泛的关注[2]。中层特征是对低层特征的一种聚集和整合,其本质是通过对低层特征的统计分布分析建立与语义之间的联系。近年来的视觉词包(BOVW)模型[3-4]就是这样一类中层特征,它无需分析场景图像中的具体目标组成,而是应用图像场景的整体统计信息,将量化后的图像低层特征视为视觉单词,通过图像的视觉单词分布来表达图像场景内容。已在图像分析和图像分类的应用中取得了巨大成功,成为了一种新的、有效的图像内容特征表示方法。

近十年来,视觉词包模型用于图像场景分类的研究层出不穷。但总体来讲,基于视觉词包模型的语义建模方法大致可划分为两大类:一类方法着重将文本分类中的思想运用于图像领域,利用概率生成模型对视觉单词进行更高层次的抽象(如视觉主

题),从而得到对场景图像(类比于文档)更贴切的语义描述[5-6];另一类方法则从图像与文本之间的区别出发,着重考虑在视觉词包模型的构建中融入空间信息,如早期的空间分布信息的利用[7-10]到后来的空间共生和上下文信息的使用[11-14]。迄今为止,国内外关于图像场景分类中视觉词包模型方法的综述性文章很少,而且国内的研究还处于起步的阶段。为了使国内同行对图像场景分类中的视觉词包模型方法有一个较为全面的了解,在参考国内外大量文献的基础上,对现有图像场景分类(本文主要指针对单一图像场景的分类)中出现的各种视觉词包模型方法从低层特征的选择与局部图像块特征的生成、视觉词典的构建、视觉词包特征的直方图表示、视觉单词优化等多方面总结和比较,同时给出了视觉词包模型性能评价的方法,介绍了几种目前常用的标准场景库,最后指出了当前研究方法中存在的问题以及该领域的发展趋势。

1　研究现状

1.1　历史与现状

早在 20 世纪 90 年代末,Joachims 就提出利用文中关键词出现的频率来表达文档的内容,并将这种表达方式用于文本分类之中[15],即词包模型(BOW)。后来,21 世纪初,一些学者基于这一特征表达思想,将词包模型引入到视觉领域,并应用于图像的纹理表示[16-17],即通过统计图像中各个不同纹理块的分布来表达图像的内容。

随后,2004 年,Csurka 等学者首次正式将词包模型用于图像场景分类的研究中[4],并提出了一个针对图像场景分类的视觉词包模型算法。通过把图像中小块纹理特征的提取推广到各种不同局部特征的提取,进而统计图像中各个图像块不同局部特征的分布信息,即把图像中的图像块对应为文本中的单词,实现了图像的视觉词包模型表示。

2005—2008 年期间,针对视觉词包模型在图像场景分类中的研究掀起一阵热潮[5,7,18-22]。不同的学者对早期的视觉词包模型进行了一系列的研究与拓展,并从图像块的划分、局部特征提取和视觉单词构造等多个阶段进行了进一步的广泛研究[23-27]。比较具有代表性的是,Sivic 等人提出的双峰单词(doublet visual words)概念,将经常成对出现的两个相邻单词定义为同一个单词,从而反映这两个单词的空间关联[19],但这种方法主要是针对图像中的目标分类,在场景分类的研究中并没有得到广泛应用。一些概率生成模型如概率潜在语义(pLSA)模型和潜在狄里克雷(LDA)模型[28]也在这一阶段广泛应用于视觉词包模型的研究中[5-6,18],这类方法通过潜在语义建立起视觉单词与图像之间的桥梁,很好地实现了对视觉单词的二次抽象,然而,由于其算法复杂度较高,计算过程也相对费时。Lazebnik 等人开创性地在视觉词包模型中加入空间金字塔匹配(SPM)核[29],提出基于空间金字塔核的词包模型,将相邻的图像块关联在一起[7]。Lazebnik 等人的这一研究成果由于其简单而有效的空间划分思路为后期不同学者关于空间信息在视觉词包模型中的应用研究提供了基础。

2009 年起,针对视觉词包模型的研究开始趋向于空间上下文或空间共生信息在视觉词包模型中的应用[11-14,30-32],期间也出现了一些有代表性的算法。Zhang 等人利用 PageRank 和 VisualRank 的思想对原有视觉单词进行筛选,得到描述性视觉单词和描述性视觉词组,共同用于图像内容表达[31-32],这种方法可以得到描述性较强的视觉词包特征,但是其描述性视觉单词和视觉词组的筛选结果受原始视觉单词质量的影响较大。Bolovinou 等人提出了基于有序空间结构关系的视觉单词,在内容表达中很好地加入上下文信息[14],但是这种方法在计算过程中要同时考虑不同范围和方向内单词的分布情况,计算量较大,并不适用于大规模场景图像分类任务。还有一些方法将图像领域中已有的空间关系分析算法引入到视觉词包模型中,如空间相关图[11],空间灰度共生矩阵(GLCM)方法[12,33],区域随机场模型[13]等,这些方法都是从已有成熟算法中吸取思路,将之拓展至视觉词包模型构建中,并取得了一定效果,但是这些方法由于复杂度较高或是效果提升不明显而没有得到广泛应用。

值得一提的是,2010 年以后,国内一些期刊和学位论文中开始逐步出现关于视觉词包模型研究的少量报导[10,34-40]。这些报导主要侧重于对视觉词包中空间分布或上下文信息的研究或是已有视觉词包算法在场景分类研究中的应用。如刘硕研提出的基于上下文语义信息的图像块视觉单词生成算法,利用视觉单词之间的语义共生概率和马尔可夫随机场理论将图像块在特征域的相似性同空间域的上下文语义共生关系有机地结合起来[34]。还有,周鸽在其

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 3, Mar. 2014

学位论文中重点对几种能够结合空间信息的算法进行了引入和改进,用于构建视觉词典,并取得了不错的效果[40]。

另外,近几年来,借助新技术、新方法对视觉词包优化的研究也成为一个新的方向。如基于流行几何学的视觉词典学习方法[41]、基于高斯混合模型的视觉词典生成方法[42]、基于 HE(hamming embedding)方法和 WGC(weak geometric consistency constraints)方法的视觉词包表达方法[43]、基于 WAFs(word activation forces)统计的视觉词包优化方法[44]、基于模糊理论的视觉词包构建方法[45]以及基于生物学驱动的视觉注意模型(visual attention model)的视觉词包模型[46-47]等。上述这些新方法都属于对视觉词包模型改进的一种新尝试,而真正要在实际应用当中得到进一步推广还需要充分的实验验证。

总的来说,国内外开展图像场景分类中的视觉词包模型的研究已有近十年的历史。整个发展过程大致可以划分为以下几个阶段:

1)早期词包模型在图像领域的直接应用阶段;

2)视觉词包模型中潜在语义信息(即对视觉单词进行二次抽象后的视觉主题信息)的研究阶段;

3)视觉词包模型中空间分布或结构信息的研究阶段;

4)视觉词包模型中上下文信息和共生信息的研究阶段;

5)视觉词典中单词语义优化及新方法引入的研究阶段。

视觉词包模型在图像场景分类研究的整个发展阶段中,所出现的研究方法种类繁多,然而,关于图像场景分类中视觉词包模型方法的综述性文章在国内外杂志上还少有报导,因此有必要对这些研究工作进行系统总结。

1.2 视觉词包模型的有关概念

为了使国内同行对视觉词包模型有一个清晰的认识,对视觉词包模型中涉及的有关概念进行了进一步的解释和说明。

1)视觉单词。类比于文档中的单词、词汇或关键词,在图像领域被称为视觉单词,可以理解为一幅图像中若干图像块的不同语义概念(如水,天空,岩石等),通常将图像块特征聚类得到的聚类中心作为视觉单词。

2)视觉词典。也称为视觉词汇表,类比于文档中由若干单词构成的词汇表,在图像领域,可以理解为视觉单词的集合。

3)视觉词包。一种中层特征表示方法,可以看做是对低层视觉特征的一种聚集和整合,通过利用图像中包含的视觉单词的统计或分布来表达图像场景内容,通常采用视觉单词频度直方图的方式表示图像特征。

4)潜在语义。类比于文档中的主题,是对单词语义的进一步概括,文档中多个单词可能同属于一个主题,在图像领域,潜在语义是对视觉单词二次抽象后的视觉主题信息(如海岸、岩石都属于自然风景主题,机场、码头都属于人工建筑主题等),是比视觉单词更高一层的中层特征。

2 研究方法分类

当前,图像场景分类中的视觉词包模型方法种类繁多,通常都涉及以下几个方面的研究内容:低层特征的选择与局部图像块特征的生成、视觉词典的构建、视觉词包特征的直方图表示、视觉单词优化等多个部分。将得到的视觉词包特征应用于监督训练的框架(如图 1 所示)中完成图像的场景分类。图 2 给出了视觉词包模型在图像场景分类中应用的整个流程。

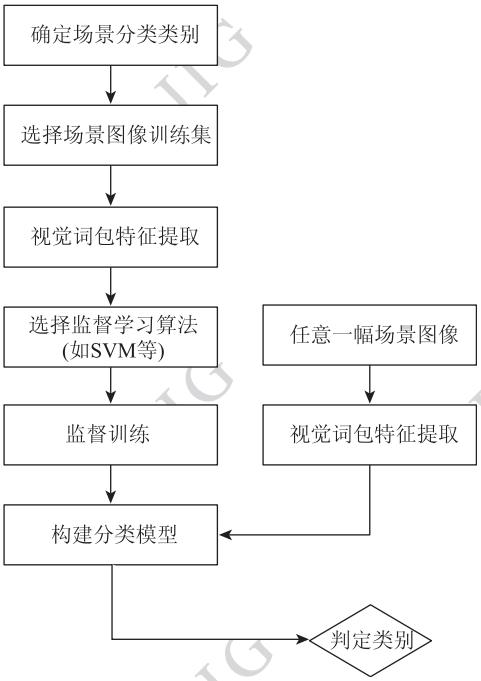


图 1 基于视觉词包特征的监督训练分类框架

Fig. 1 Bag-of-visual words based supervised training classification framework

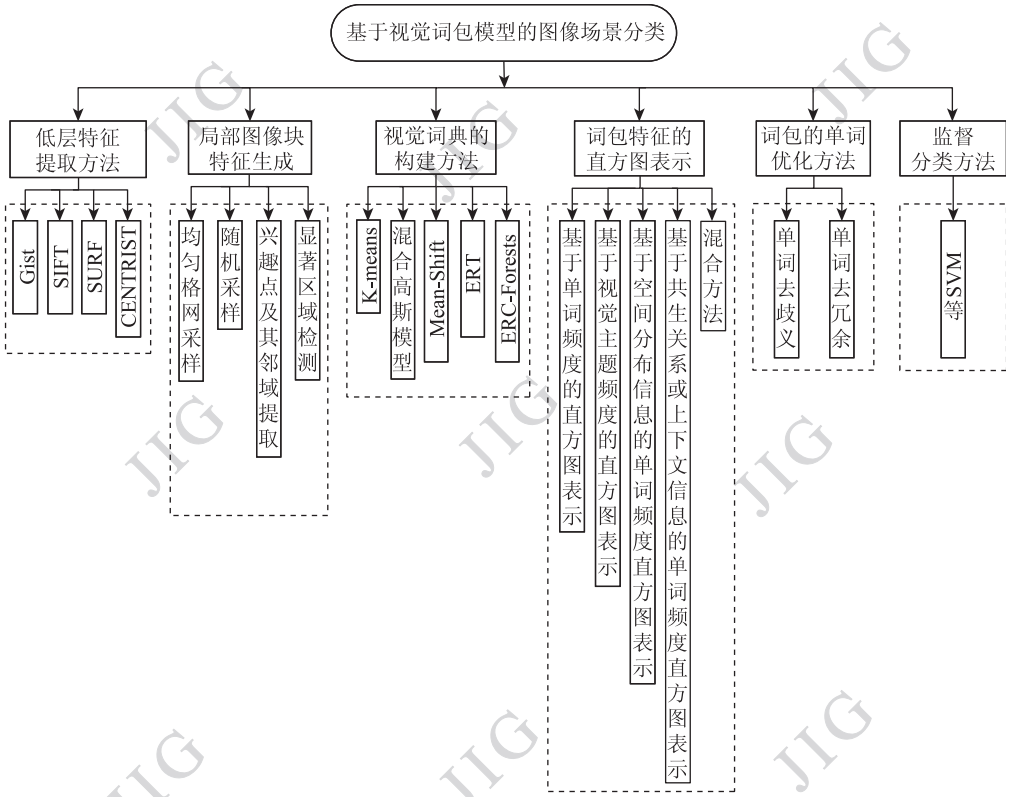


图 2　基于视觉词包模型的图像场景分类方法

Fig. 2　Bag-of-visual-words model based image scene classification methods

2.1　低层特征的选择与局部图像块特征的生成

在视觉词包模型的研究中，低层视觉特征的选取对于视觉词典的构建具有很大影响。因此，低层特征的选择十分重要，往往需要考虑到特征的不变特性，如旋转、尺度不变性等。在不变性方面表现比较好的特征有：Gist[48]、SIFT[49]、SURF[50]、CENTRIST[51]等。其中，由于 SIFT 特征具有很好的几何变化、光照、遮挡和视角变化的鲁棒性，成为图像视觉词包模型中最常见的低层特征提取算法。

在选定了低层特征提取算法的基础上需要对场景图像按照一定的采样方法生成采样点，进而计算以各采样点为中心的邻域图像块的低层特征作为局部图像块特征，进而基于这些局部图像块特征构建视觉词典。目前，局部图像块特征生成中的采样方法多种多样，总的来说，可以归纳为以下几种（以 SIFT 特征提取为例）：

1）均匀网格采样方法。在给定图像中均匀采样图像块，每个图像块的大小为 $N \times N$，图像块之间的间隔为 $M \times M$（当 $M < N$ 时，采样的网格会产生重叠），以各图像块的中心作为采样点，计算 SIFT 特征。

2）随机采样方法。在给定的图像中随机采样预定个数的图像块（大小为 $N \times N$），然后以各图像块的中心点作为采样点，计算 SIFT 特征。

3）兴趣点及其邻域提取。由兴趣点检测算子提取图像中所有的兴趣点，将每个点的邻域作为图像块，并以各兴趣点作为采样点计算 SIFT 特征。

4）显著区域检测。根据 Itti 等人提出的显著性模型[52]，计算视觉显著区域，作为图像块，并以图像块中心点作为采样点，计算 SIFT 特征。

在场景理解任务中，Li 和 Boash 等人指出，无论何种兴趣点检测方法都不能得到图像内足够的信息，而必须选取图像内所有的小块表示图像场景内容[5-6]。因此，在场景分类过程中，利用均匀网格采样来提取的图像块的方法得到了广泛的应用。

2.2　视觉词典的构建方法

目前，生成视觉词典的方法主要有两种：一种是 Vogel 等人所采用的人工标注的方法[26]；另一种是无监督聚类算法[23]。由于人工标注存在着工作量巨大和主观性强等问题，使得无监督聚类算法成为目前生成视觉单词最主要的方法。在无监督聚类方法中，使用最为广泛的方法是 K 均值聚类方法，通

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 3, Mar. 2014

过对图像集中所有提取的图像块的局部特征进行聚类,将每个聚类中心作为视觉词典中的一个单词。通常情况下,聚类中心个数 K 的选择将影响视觉词典的性能。一般而言,当 K 较小时,算法得到少量的视觉单词,可能导致两个差异较大的图像块被分配到同一个集群(即指派为同一个视觉单词);当 K 增大时,算法产生更多的视觉单词,从而导致一些相似的图像块被"过分割"为两个集群,降低视觉单词的泛化能力,并且增大计算量。因此,在图像的视觉词典的构建过程中,需要平衡视觉单词的描述能力和泛化能力,确定一个合适的词典大小[37]。

除了基于 K 均值聚类算法的视觉词典构建方法外,不少研究人员也尝试利用多种不同的学习方法来构建视觉词典。如利用混合高斯模型[53]、Mean-Shift 方法[54]、完全随机树(ERT)[55-56]和完全随机聚类森林(ERC-Forests)[56]等方法对图像块进行聚类,构造视觉词典。但是这些方法过于复杂,并没有得到广泛的应用。

2.3 视觉词包特征的表示方法

通常采用直方图表示。视觉词包特征的直方图表示是视觉词包模型在图像场景分类过程中的核心内容。从方法提出起,出现了多种视觉词包特征的直方图表示方法。其中,一类方法从文本分类领域中借鉴相应的思路和方法展开研究,如基于视觉单词频度的直方图表示、基于对视觉单词二次抽象的视觉主题频度的直方图表示;另一类方法,则从图像与文本之间的差异出发,对视觉词包方法进行了空间上的延伸和拓展,如基于空间分布信息的视觉单词频度的直方图表示、基于共生关系或空间上下文信息的视觉单词频度的直方图表示。

2.3.1 借鉴文本分类思路的词包特征直方图表示

1)基于视觉单词频度的直方图表示。这类表达方法主要见于视觉词包模型提出的早期,通过计算训练图像中图像块的特征与视觉词典中的每个视觉单词所对应的欧氏距离,记录与图像块特征距离最近的某个视觉单词。将图像中的图像块重复上面过程,于是形成一组视觉单词频度统计直方图,利用这组统计直方图代表该幅图像的特征。早期的这种经典视觉词包模型实现简单,但是仅仅依据图像块与视觉单词在特征空间中的欧氏距离,并没有考虑到图像中的空间信息以及图像块所对应的视觉单词在语义空间中是否与其周围图像块具有共生性。

2)基于视觉主题频度的直方图表示。视觉主题是对视觉单词的二次抽象。这类方法主要从主题模型(topic model)的角度出发将图像表达为一个包含多个隐藏主题的混合分布,而这些隐藏主题则是对视觉单词的二次抽象,比视觉单词具有与图像内容更加相关和贴切的语义含义。比较经典方法有:基于概率潜在语义模型的视觉词包特征的直方图表示[6,57]和基于潜在狄里克雷模型的视觉词包特征的直方图表示等[5]。其中,基于概率潜在语义模型的视觉词包特征的直方图表示通过潜在语义分析建立图像、主题、视觉单词三者之间的关系,利用视觉主题表达图像内容,降低了特征的维数,但是主题的概率是从实际观察变量中得到的离散值,没有对其建立通用的概率模型,模型参数个数随着图像个数的线性增长可能会产生过拟合;基于潜在狄里克雷模型则克服了概率潜在语义模型中的缺点,将每幅图像的概率视为潜在主题中随机出现视觉单词概率的混合模型,但是这也在很大程度上提高了模型的复杂度。

2.3.2 融入空间信息的视觉单词频度的直方图表示

1)基于空间分布信息的视觉单词频度的直方图表示。这类方法主要是通过在图像空间或是特征空间进行不同尺度的分层处理来实现对空间分布信息的利用。最具代表性的方法是基于空间金字塔匹配的视觉词包特征的直方图表示[7],将图像不断分割成越来越细的子区域,构造相应的空间金字塔,并对每个子区域提取相应的直方图特征,并应用加权方式直接连接这些直方图,构造成一个"更大"的直方图表示,从而得到基于空间金字塔匹配核的词包表示。然而,该方法忽略了特征之间的空间共生关系,且各层级的权重也需要经验确定,因此,在此基础上,又延伸出来很多其他方法。如 Zhou 等人[9]通过构造多分辨率图像,并对每个分辨率下的图像进行一定规则的分块来提取特征,以聚类构造视觉词典,然后在表达图像特征的时候,将各分辨率下的图像进行水平和垂直方向的分块处理,通过统计每个子块对应的视觉单词得到最终的视觉单词频度直方图,但是当图像分辨率数目较多时,这种方法在各尺度权重的确定上仍然存在较大困难。Qin 等人[8]除了对图像进行不同尺度的表示,每个尺度又考虑了邻近和粗尺度的上下文信息,同时还在不同的尺度下构造不同的视觉词典,最后通过将所有各类场景中每一个尺度下的词典连接构建最终的视觉词典,

用于最终场景图像特征的视觉单词频度表示。该方法在上下文信息的构建中对局部信息采用线性组合的方法，操作简单，但在信息组合过程中对于图像块特征的考虑不足。王宇新等人[10]则对视觉词典进行了更细致地划分，将图像进行不同层次的空间划分，针对对应空间中各对应子区域形成该区域的视觉词典，并将所有空间各子区域的视觉词典连接成一个全局特征向量进行图像场景内容表达。这种方法所提出的图像空间划分思路取得了较好的结果，但是图像是否需要动态划分或重叠划分，以及划分为多少级才能最大程度地提高分类准确率，尚需进一步的实验和理论证明。

2）基于共生关系或空间上下文信息的视觉单词频度的直方图表示。这类方法主要考虑加入图像中视觉单词的空间共生关系以及一些上下文信息来进行场景内容的视觉单词频度表达。比较传统的方法如视觉单词成对体[21,31,58-59]或三联体[62]（通常也被称为视觉词组或视觉短语）。Zheng 等人[11]基于颜色相关图[63]的思想，提出空间相关图的概念，并将之应用于视觉词包特征表达，在加入了几何信息的同时提升了场景分类的性能，该方法很好地利用了视觉单词之间的局部几何关系信息，但是相关图的特征表达方式在一定程度上增加了存储和计算的需求量。除此之外，还有其他的一些结合视觉单词共生关系的视觉词包表达方法，如基于区域条件随机场模型考虑一定范围内不同视觉单词间的联系[13]，利用空间灰度共生矩阵思想发掘视觉单词之间的共生关系[12]等。

另外，一些新出现的方法开始以组群（group）的方式考虑视觉单词[60,64]，以此来加入局部图像特征的空间上下文信息，同时也更有效地获取到视觉单词之间的空间结构信息。然而，在视觉词典生成过程中引入的量化误差则可能降低视觉单词组合的匹配精度。针对这一问题，Zhang 等人[64]提出了一种上下文视觉词典，首先将多个局部特征以组群的方式聚集在一起，然后将这些局部特征组群量化为视觉单词，从而使每一个组群中的空间信息在生成的局部特征中很好地保留下来。与此同时，由于没有视觉单词组合的操作步骤，量化误差被很好地抑制。

再者，有序的空间结构关系也能够很好地挖掘图像场景内容表达中的视觉单词的上下文信息。如 Bolovinou 等人[14]考虑了不同方向不同范围内的视觉单词出现情况，提出构建具有有序空间结构关系的视觉词典从而在内容表达中很好地加入了上下文信息，但是在单词频度统计中要同时考虑到距离和方向两个因素，因而计算量较大。

2.3.3　混合方法

综合上述两大类方法的特点，研究人员结合多类方法对图像场景进行视觉词包特征的直方图表示，其中比较经典的算法如 Bosch 等人[61]所提出的 SP-pLSA 方法。该算法将对视觉单词二次抽象的视觉主题信息与场景图像特有的空间信息相结合，共同来构建视觉词包模型。其中视觉主题信息采用概率潜在语义模型学习无标记的训练图像的潜在主题及其分布情况，并利用其在测试图像中的分布作为特征向量用于监督判别式分类器。同时，为了加入图像场景的空间信息，该算法又尝试将多种空间信息特征表达方式与概率潜在语义模型相结合，最后得到效果最好的融合空间金字塔匹配和概率潜在语义模型的视觉词包构建方法用于图像场景特征的直方图表示。然而，该方法对于视觉单词之间的空间共生关系和上下文信息考虑还不够充分。其他的混合方法如 Gong 等人[65]提出的将概率潜在语义模型与条件随机场模型相结合，利用概率潜在语义模型对图像中每两个视觉单词建立共生关系，再利用条件随机场模型为他们的空间位置信息建模，并给出概率模型，用于解释场景类别的结构信息，但是，由于需要计算概率模型，在一定程度上提高了模型的复杂度。

2.4　视觉词包的单词优化方法

视觉词包的单词优化方法主要针对视觉词包构建过程中存在的单词语义歧义性、模糊性、单词冗余等问题[45]进行优化。

视觉单词的歧义性和模糊性会造成同一个视觉单词对于不同类别的贡献以及不同的视觉单词对于同一个类别的贡献是相同的或相似的，从而对于图像场景分类效果产生很大影响。一些方法通过剔除出现频率最高和最低的视觉单词来消除模糊性问题[3]；还有一些方法则将一幅图像的特征分配给特征空间中多个邻近的视觉单词[24]。相类似地，也有研究人员使用概率视觉单词投票策略，即给定视觉单词，每一个图像特征会根据图像特征的后验概率为多个视觉单词投票[53,66]。由于是考虑的是多个视觉单词，所以这些方法能够很好地处理视觉单词的歧义性和模糊性问题，并且能够在一定程度上提高场景分类的精度。

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 3, Mar. 2014

另一方面,针对视觉单词数量大的问题,出现了许多视觉单词去冗余的方法,如:主成分分析法[67]、WAFs(word activation forces)统计法[44]等。它们都是通过降低视觉词包的冗余度来提升图像内容的表达能力。为了获取更有区分度的视觉词典,共生矩阵也被用来估计各场景类别中每一个视觉单词的重要性[42,68],从而根据重要性有选择地提炼视觉单词。

3 视觉词包模型性能评价方法

目前,在视觉词包模型的研究中,通常会借助某一监督学习分类器,采用准确率这一指标来衡量不同模型的性能好坏。具体来讲,以某一场景图像数据集作为评价样本,通过对数据集进行一定比例的随机划分得到训练样本和测试样本,其中训练样本用于视觉词典的构建,完成对场景图像的视觉词包特征表达,以及对分类器分类模型的学习和构建,而测试样本则主要依据分类器的分类结果,用于准确率的计算,评价模型的性能。通常情况下,为了得到令人信服的性能评价结果,会多次重复整个实验过程(包括训练/测试样本的划分,视觉词典的构建,图像视觉词包特征表达等内容),最终以平均准确率和标准偏差作为最终的性能评价结果。平均准确率数值越高说明视觉词包模型在图像场景内容表达方面的性能越好,标准偏差在一定程度上反映了模型的稳定性,标准偏差越小说明模型受样本变化的影响越小。

另外,为了得到进一步细化的性能评价结果,通常还会采用混淆矩阵的方法来评价视觉词包模型在不同类别的场景图像内容表达中的性能优劣。其中,位于混淆矩阵对角线上的元素数值反映了视觉词包模型对于每个类别的场景图像分类的准确率,准确率越高则说明其性能越好。

4 常用标准场景库介绍

为了更方便地进行算法比较,对目前常用的几个标准场景库进行了汇总介绍,如表 1 所示。

表 1 常用的标准场景库介绍

Table 1 Description of the widely used standard scene datasets

场景库名称	建立者及年份	类别数	描述	当前最高精度/%
OT	Oliva 和 Torralba,2001[48]	8	彩色图像,包含 2 688 幅图像,平均尺寸为 250×250 像素	92.79[69]
FP	Li 和 Perona,2005[5]	13	灰度图像,其中包含 OT 场景库的 8 类图像,平均尺寸为 250×300 像素	86.10[9]
LSP	Lazebnik 等人,2006[7]	15	灰度图像,包含 4 486 幅图像,其中包含 FP 场景库中的 13 类图像,每类包含 200～400 幅图像,平均尺寸为 300×250 像素	86.10[70]
LF	Li 和 Li,2007[71]	8	包含 8 类运动场景图像,每类包含 137～250 幅图像	85.10[9]
QT	Quattoni 和 Torralba,2009[72]	67	包含 15 620 幅彩色图像	46.50[9]
MSRC-18	Shotton 等人,2007[73]	18	该场景库提供了图像中目标的标注信息	75.24[14]

5 结 语

由以上综述可以看到,图像场景分类中视觉词包模型方法的研究作为计算机视觉领域方兴未艾的热点研究领域,在国内外研究中取得了不少进展,在计算机视觉领域的研究也不再局限于直接应用模型描述图像内容,而是更多地考虑图像与文本的差异,从图像块提取、视觉单词的构造和映射,以及图像块的独立性假设等方面研究提高词包模型对图像场景的描述能力。虽然视觉词包模型在图像场景分类的应用中还存在很多亟需解决的问题,但是这丝毫不能掩盖其研究的重要意义。

总结图像场景分类中视觉词包模型方法的研究,本文认为存在以下亟待解决的问题:

1)当训练得到的视觉词包应用于与训练样本有很大不同的样本时,视觉词包的性能会受到很大影响,而基于新的样本集训练新的视觉词包又是十

分费时费力的。

2)尽管目前已有许多关于视觉词包模型的研究,但是对于视觉词典的大小的确定仍然缺乏足够的理论指导。

3)对于视觉单词与语义含义之间的关系还没有被充分挖掘,同义词和多义词是不能回避的问题,在这方面的研究还很不充分。

4)对于特殊领域特定应用的研究还比较少,如高分辨率遥感图像中人工场景或土地利用的分类问题,需要研究如何将视觉词包模型与高分辨率遥感图像场景分类或土地利用分类有机结合起来。

针对上述问题,目前视觉词包模型方法用于图像场景分类的研究主要集中在如下方面:

1)构建通用的自适应的视觉词包用于不同的样本集。

2)根据具体问题自动选择最优化的视觉词包大小。

3)在图像的词包表示中加入更加丰富的空间分布和上下文信息,发掘视觉单词的潜在语义信息。

4)对图像视觉语法进行研究,使其作为图像模型或主题模型服务于图像理解。

5)研究特殊领域图像的场景分类问题,如高分辨率遥感图像中城市功能区的识别等。

6)研究新的具有良好性能的低层特征提取算法来构建高层的视觉词包。

参考文献(References)

[1] Gao J, Xie S. The Theory and Method of Image Understanding [M]. Beijing: Science Press, 2009. [高隽, 谢昭. 图像理解理论与方法 [M]. 北京: 科学出版社, 2009.]

[2] Datta R, Joshi D, Li J, et al. Image retrieval: ideas, influences, and trends of the new age[J]. ACM Computing Surveys, 2008, 40(2): 1-60.

[3] Sivic J, Zisserman A. Video google: a text retrieval approach to object matching in videos[C]//Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France: IEEE, 2003: 1470-1477.

[4] Csurka G, Dance C R, Fan L, et al. Visual categorization with bags of keypoints[C]//Proceedings of Workshop on Statistical Learning in Computer Vision. Prague, Czech Republic: Springer, 2004: 1-22.

[5] Li F F, Perona P. A bayesian hierarchical model for learning natural scene categories[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, United States: IEEE, 2005: 524-531.

[6] Bosch A, Zisserman A, Muñoz X. Scene classification via pLSA [C]//Proceedings of the 9th European Conference on Computer Vision. Graz, Austria: Springer, 2006: 517-530.

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

[8] Qin J Z, Yung N H C. Scene categorization via contextual visual words[J]. Pattern Recognition, 2010, 43(5): 1874-1888.

[9] Zhou L, Zhou Z, Hu D. Scene classification using a multi-resolution bag-of-features model [J]. Pattern Recognition, 2013, 46(1): 424-433.

[10] Wang Y X, Guo H, He C Q, et al. Bag of spatial visual words model for scene classification [J]. Computer Science, 2011, 38(8): 265-268. [王宇新, 郭禾, 何昌钦, 等. 用于图像场景分类的空间视觉词袋模型[J]. 计算机科学, 2011, 38(8): 265-268.]

[11] Zheng Y, Lu H, Jin C, et al. Incorporating spatial correlogram into bag-of-features model for scene categorization[C]//Proceedings of the 9th Asian Conference on Computer Vision. Xi'an, China: Springer, 2010: 333-342.

[12] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification[C]//Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. San Jose, CA, United States: ACM, 2010: 270-279.

[13] Liu S Y, Xu D, Feng S H. Region contextual visual words for scene categorization [J]. Expert Systems with Applications, 2011, 38(9): 11591-11597.

[14] Bolovinou A, Pratikakis I, Perantonis S. Bag of spatio-visual words for context inference in scene classification[J]. Pattern Recognition, 2013, 46(3): 1039-1053.

[15] Joachims T. Text categorization with support vector machines: Learning with many relevant features[J]. Lecture Notes in Computer Science, 1998, 1398: 137-142.

[16] Leung T, Malik J. Representing and recognizing the visual appearance of materials using three-dimensional textons[J]. International Journal of Computer Vision, 2001, 43(1): 29-44.

[17] Cula O G, Dana K J. Compact representation of bidirectional texture functions[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, United States: IEEE, 2001: 1041-1047.

[18] Fergus R, Li F F, Perona P, et al. Learning object categories from Google's image search[C]//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005: 1816-1823.

[19] Sivic J, Russell B C, Efros A A, et al. Discovering objects and their location in images[C]//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005: 370-377.

[20] Winn J, Criminisi A, Minka T. Object categorization by learned universal visual dictionary[C]//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005: 1800-1807.

[21] Yuan J, Wu Y, Yang M. Discovery of collocation patterns: from visual words to visual phrases[C]//Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, United States: IEEE, 2007: 1-8.

[22] Hotta K. Object categorization based on kernel principal component analysis of visual words[C]//Proceedings of the IEEE Workshop on Applications of Computer Vision. Copper Mountain, CO, United States: IEEE, 2008: 1-8.

[23] Nowak E, Jurie F, Triggs B. Sampling strategies for bag-of-features image classification[C]//Proceedings of the 9th European Conference on Computer Vision. Graz, Austria: Springer Berlin Heidelberg, 2006: 490-503.

[24] Jiang Y G, Ngo C W, Yang J. Towards optimal bag-of-features for object categorization and semantic video retrieval[C]//Proceedings of the 6th ACM International Conference on Image and Video Retrieval. Amsterdam, Netherlands: ACM, 2007: 494-501.

[25] Quelhas P, Monay F, Odobez J M, et al. A thousand words in a scene[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(9): 1575-1589.

[26] Vogel J, Schiele B. Semantic modeling of natural scenes for content-based image retrieval[J]. International Journal of Computer Vision, 2007, 72(2): 133-157.

[27] Zeng P, Wen J, Wu L D. Scene classification using low-level feature and intermediate feature[C]//Proceedings of the MIPPR 2007: Pattern Recognition and Computer Vision. Wuhan, China: SPIE, 2007.

[28] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(4-5): 993-1022.

[29] Grauman K, Darrell T. The pyramid match kernel: discriminative classification with sets of image features[C]//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005: 1458-1465.

[30] Tian Q, Zhang S, Zhou W, et al. Building descriptive and discriminative visual codebook for large-scale image applications[J]. Multimedia Tools and Applications, 2011, 51(2): 441-477.

[31] Zhang S L, Tian Q, Hua G, et al. Descriptive visual words and visual phrases for image applications[C]//The 17th ACM International Conference on Multimedia. Beijing, China: ACM, 2009.

[32] Zhang S L, Tian Q, Hua G, et al. Generating descriptive visual words and visual phrases for large-scale image applications[J]. IEEE Transactions on Image Processing, 2011, 20(9): 2664-2677.

[33] Haralick R M, Shanmugam K, Dinstein. I. Texture features for image classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1973, 3: 610-621.

[34] Liu S Y, Xu D, Feng S H, et al. A novel visual words definition algorithm of image patch based on contextual semantic information [J]. Acta Electronica Sinica, 2010, 38(5): 1156-1161. [刘硕研, 须德, 冯松鹤, 等. 一种基于上下文语义信息的图像块视觉单词生成算法[J]. 电子学报, 2010, 38(5): 1156-1161.]

[35] Wang Y S, Gao W. Kernel-based image classification using the context of visual words[J]. Journal of Image and Graphics, 2010, 15(4): 607-616. [王宇石, 高文. 用基于视觉单词上下文的核函数对图像分类[J]. 中国图象图形学报, 2010, 15(4): 607-616.]

[36] Liu Y W, Huo H, Fang T. Visual words ambiguity analysis in BOW model[J]. Computer Engineering, 2011, 37(19): 204-206, 209. [刘扬闻, 霍宏, 方涛. 词包模型中视觉单词歧义性分析[J]. 计算机工程, 2011, 37(19): 204-206, 209.]

[37] Xu S. Topic model on image classification and their applications in high spatial resolution remote sensing image[D]. Shanghai: Shanghai Jiaotong University, 2012. [徐盛. 基于主题模型的高空间分辨率遥感影像分类研究[D]. 上海: 上海交通大学, 2012.]

[38] Tang Y J. Research on semantic topic model based image scene classification[D]. Beijing: Beijing Jiaotong University, 2010. [唐颖军. 基于语义主题模型的图像场景分类研究[D]. 北京: 北京交通大学, 2010.]

[39] Xie W J. Research of middle-level semantic based image scene classification algorithm[D]. Beijing: Beijing Jiaotong University, 2011. [解文杰. 基于中层语义表示的图像场景分类研究[D]. 北京: 北京交通大学, 2011.]

[40] Zhou G. Bag-of-visual words for image categorization[D]. Suzhou: Soochow University, 2011. [周鸽. 基于"词袋"模型的图像分类系统[D]. 苏州: 苏州大学, 2011.]

[41] Tian X, Lu Y. Discriminative codebook learning for Web image search[J]. Signal Processing, 2013, 93(8): 2284-2292.

[42] Liu G. Improved bags-of-words algorithm for scene recognition[J]. Journal of Computational Information Systems, 2010, 6(14): 4933-4940.

[43] Jégou H, Douze M, Schmid C. Improving bag-of-features for large scale image search[J]. International Journal of Computer Vision, 2010, 87(3): 316-336.

[44] Li Q, Zhang H G, Guo J, et al. Improving bag-of-words scheme for scene categorization[J]. The Journal of China Universities of Posts and Telecommunications, 2012, 19(S2): 166-171.

[45] Van Gemert J C, Veenman C J, Smeulders A W M, et al. Visual word ambiguity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(7): 1271-1283.

[46] Xu S, Fang T, Huo H, et al. A novel method of aerial image classification based on attention-based local descriptors[J]. Procedia Earth and Planetary Science, 2009, 1(1): 1133-1139.

[47] Zhang J, Sui L, Zhuo L, et al. An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain[J]. Neurocomputing, 2013, 110: 145-152.

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

[49] Lowe D. Distinctive image features form scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 20(2): 91-110.

[50] Thomee B, Bakker E M, Lew M S. TOP-SURF: a visual words toolkit[C]//Proceedings of the 18th ACM International Conference on Multimedia. Firenze, Italy: ACM, 2010.

[51] 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.

[52] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254-1259.

[53] Perronnin F. Universal and adapted vocabularies for generic visual categorization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(7): 1243-1256.

[54] Jurie F, Triggs B. Creating efficient codebooks for visual recognition[C]//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005: 604-610.

[55] Maree R, Geurts P, Piater J, et al. Random subwindows for robust image classification[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, United States: IEEE, 2005: 34-40.

[56] Moosmann F, Triggs B, Jurie F. Fast discriminative visual codebooks using randomized clustering forests[C]//Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Vancouver, BC, Canada: NIPS, 2007.

[57] Saghafi B, Farahzadeh E, Rajan D, et al. Embedding visual words into concept space for action and scene recognition[C]//Proceedings of the British Machine Vision Conference. Aberystwyth, UK: BMVA Press, 2010: 15. 1-15. 11.

[58] Setia L, Teynor A, Halawani A, et al. Grayscale medical image annotation using local relational features[J]. Pattern Recognition Letters, 2008, 29(15): 2039-2045.

[59] Savarese S, Winn J, Criminisi A. Discriminative object class models of appearance and shape by correlatons[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, United States: IEEE, 2006, 2033-2040.

[60] Yu L, Liu J, Xu C. Descriptive local feature groups for image classification[C]//Proceedings of the 18th IEEE International Conference on Image Processing. Brussels, Belgium: IEEE, 2011, 2501-2504.

[61] Bosch A, Zisserman A, Muñoz X. Scene classification using a hybrid generative/discriminative approach[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(4): 712-727.

[62] Haibin L, Soatto S. Proximity distribution kernels for geometric context in category recognition[C]//Proceedings of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007.

[63] Jing H, Kumar S R, Mitra M, et al. Image indexing using color correlograms[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Rio de Janeiro, Brazil: IEEE, 1997: 762-768.

[64] Zhang S, Huang Q, Hua G, et al. Building contextual visual vocabulary for large-scale image applications[C]//Proceedings of the 18th ACM International Conference on Multimedia. Firenze, Italy: ACM, 2010: 501-510.

[65] Wang Y, Gong S, Mary Q. Conditional random field for natural scene categorization [C]//Proceedings of the British Machine Vision Conference. Warwick: BMVA Press, 2007: 1-10.

[66] Agarwal A, Triggs B. Multilevel image coding with hyperfeatures[J]. International Journal of Computer Vision, 2008, 78(1): 15-27.

[67] Okumura S, Maeda N, Nakata K, et al. Visual categorization method with a bag of PCA packed keypoints[C]//Proceedings of the 4th International Congress on Image and Signal Processing. Shanghai, China: IEEE, 2011: 950-953.

[68] Hao J, Jie X. Improved bags-of-words algorithm for scene recognition[C]//Proceedings of the 2nd International Conference on Signal Processing Systems. Dalian, China: IEEE, 2010: V2279-V2282.

[69] Perina A, Cristani M, Castellani U, et al. A hybrid generative/discriminative classification framework based on free-energy terms [C]//Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009: 2058-2065.

[70] Nakayama H, Harada T, Kuniyoshi Y. Global gaussian approach for scene categorization using information geometry[C]//Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, United States: IEEE, 2010: 2336-2343.

[71] Li L J, Li F F. What, where and who? Classifying events by scene and object recognition[C]//Proceedings of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007: 1-8.

[72] Quattoni A, Torralba A. Recognizing indoor scenes[C]//Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Miami, FL, United States: IEEE, 2009: 413-420.

[73] Shotton J, Winn J, Rother C, et al. Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context[J]. International Journal of Computer Vision, 2009, 81(1): 2-23.