文章编号:1001 - 9081(2008)06 - 1537 - 03

基于分块潜在语义的场景分类方法

曾 璞,吴玲达,文 军
(国防科学技术大学 信息系统与管理学院, 长沙 410073)
(puzeng_nudt@ nudt. edu. cn)

摘 要:提出了一种基于分块潜在语义的场景分类方法。该方法首先对图像进行均匀分块并使用分块内视觉词汇的出现频率来描述每一个分块,然后利用概率潜在语义分析(PLSA)方法从图像的分块集合中发现潜在语义模型,最后利用该模型提取出潜在语义在图像分块中的出现情况来进行场景分类。在 13 类场景图像上的实验表明,与其他方法相比,该方法具有更高的分类准确率。

关键词:场景分类;分块潜在语义;视觉词汇;局部不变特征;概率潜在语义分析
中图分类号: TP391. 41 **文献标志码**:A

Scene classification based on block latent semantic

ZENG Pu, WU Ling-da, WEN Jun
(*School of Information System and Management, National University of Defense Technology, Changsha Hunan* 410073, *China*)

Abstract: A novel scene classification method was presented based on block latent semantic. The image blocks were first extracted on a regular grid and the visual words in blocks were used to describe every block, and then block latent semantic models were achieved by using Probabilistic Latent Semantic Analysis (PLSA). The latent semantic model was used to find the latent semantic in image block and their spatial distribute in image. Finally, this feature was used to construct a SVM model to classify scene. Experimental results show that this method has satisfactory classification performances on a large set of 13 categories of complex scenes.

Key words: scene classification; block latent semantic; visual word; local invariant feature; Probabilistic Latent Semantic Analysis (PLSA)

0 引言

面对数量巨大的图像数据,传统的依靠人工来对图像进行分类与标注的管理方式因为需要耗费大量的人力资源而变得不可行。因此,如何利用计算机来自动将图像按照人们理解的方式分类到不同的语义类别就成为其中的一个关键问题。在人们对图像理解的众多语义内容中,图像的场景类别不仅包含了人们对一幅图像的总体认识,而且还提供了图像中对象出现的上下文环境,为进一步识别出图像中的对象提供了基础。因此,图像场景分类就成为当前计算机视觉和多媒体信息管理领域的热点问题。

根据描述图像方式的不同,当前的场景分类方法可以分为基于底层特征和基于中间语义特征两大类。使用图像底层特征来对场景进行分类的问题已经在图像和视频检索领域研究了多年。这些工作通常使用色彩、纹理和形状等图像底层特征来直接与监督学习方法结合,从而将图像分类到不同的语义类别中[1-3],如室内、室外、城市、乡村、日落、森林等。近年来,为了克服图像底层视觉特征与高层语义之间的语义鸿沟,使用中间语义特征来对场景建模的方法得到了广泛的关注。文献[4]中使用一组视觉感知属性(自然度、宽阔度、粗糙度、伸展度和险峻度)来描述场景的主要空间结构。文献[5]中首先定义一组局部语义概念,然后通过训练样本来生成局部语义概念模型,最后使用这些语义概念模型计算相应局部语义概念在图像中的出现频率来进行场景分类。在这类方法中,中间语义特征的生成往往需要大量的手工标注样本。为了减少生成中间语义特征所需的样本,文本分析中的主题模型被用于图像场景分类[6-8]。这些方法首先将图像的局部不变特征聚类为一组视觉词汇,然后用词袋(Bag Of Word, BOW)的方式来表示图像,最后用概率潜在语义分析(Probabilistic Latent Semantic Analysis, PLSA)[9]或者 LDA(Latent Dirichlet Allocation)[10]等主题分析模型来找出图像最可能属于的主题,从而完成图像的场景分类。然而,主题分析模型往往直接根据图像中视觉词汇出现的总体情况来进行场景分类,并没有考虑到人们在进行场景分类时,图像中的区域语义及其空间分布往往起着十分重要的作用。例如,一个海滩的场景往往是由在图像上部的天空、中间的海洋和下部的沙滩组成,通过分析图像在相应位置是否包含特定区域语义就能够有效地对图像场景进行分类。

考虑到现有方法中存在的这些不足,本文提出了一种基于分块潜在语义的场景分类新方法。该方法首先将训练集中图像进行均匀分块得到图像分块集合,并用分块中的视觉词汇出现的频率来描述分块,然后采用概率潜在语义分析对图像分块集合进行主题发现从而构建潜在语义模型。最后将图像各分块中潜在语义的出现情况描述为图像的分块潜在语义特征,并构建 SVM 分类器来进行场景分类。在 13 类场景图像集上的实验表明,本文方法与其他方法相比具有更高的分

收稿日期:2007 - 12 - 05;**修回日期**:2008 - 02 - 01。
基金项目:国家自然科学基金资助项目(60473117);国家 863 计划项目(2006AA01Z319)。
作者简介:曾璞(1980 -),男,湖南桃江人,博士研究生,主要研究方向:多媒体信息系统、计算机视觉; 吴玲达(1962 -),女,上海人,教授,博士生导师,主要研究方向:多媒体信息系统、虚拟现实; 文军(1976 -),男,湖南邵阳人,博士研究生,主要研究方向:多媒体信息系统。

类准确性。

1 PLSA方法

自然语言处理(Natural Language Processing, NLP)的很多应用都需要探究隐藏在字、词背后的涵义,简单的字面匹配绝难奏效,关键在于同义词和一词多义的把握。潜在语义分析(LSA)为此提供了部分解决问题的方法,即利用奇异值分解(Singular Value Decomposition, SVD)将高维度的词汇-文档共现矩阵映射到低维度的潜在语义空间,使得表面毫不相关的词体现出深层次的联系。概率潜在语义分析作为潜在语义分析(Latent Semantic Analysis, LSA)的变种,拥有更坚实的数学基础及易于利用的数据生成模型,且已被证实能够为信息提取提供更好的词汇匹配。

给定一个文档集合 $D = \{d_1, d_2, \cdots, d_M\}$ 和一个词集合 $W = \{w_1, w_2, \cdots, w_N\}$,以及一个文档和词的共现频率矩阵 $N \equiv (n_{ij})$,$n(d_i, w_j)$ 表示词 w_j 在文档 d_j 中出现的频率。使用 $Z = \{z_1, z_2, \cdots, z_K\}$ 表示潜在语义的集合,K 为人工指定的一个常数。概率潜在语义分析假设"文档 - 词" 对之间是条件独立的,并且潜在语义在文档或词上分布也是条件独立的。在上面假设的前提下,可使用式(1) 来表示"文档 - 词" 的条件概率:

$$P(d_i, w_j) = P(d_i) \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i) \tag{1}$$

其中,$P(w_j \mid z_k)$ 为潜在语义在词上的分布概率,也可以解释为词对潜在语义的贡献度。$P(z_k \mid d_i)$ 表示文档中的潜在语义分布概率,也可以解释为文档中具有相应潜在语义的概率。

概率潜在语义分析使用期望最大(Expectation Maximization, EM)算法来求取式(2)中对数似然函数的极大值,从而拟合潜在语义模型,得到 $P(w_j \mid z_k)$ 和 $P(z_k \mid d_i)$ 的分布:

$$\sum_{d \in D} \sum_{w \in W} n(d_i, w_j) \log P(d_i, w_j) \tag{2}$$

2 分块潜在语义提取

2.1 视觉词汇表的生成

要使用文本分析中的PLSA等主题模型来对图像进行分析,首先需要做的就是在图像中构建出与文本中的词相对应的视觉词汇。文献[7]提供了一种在图像中生成视觉词汇的方法。该方法首先用DoG(Difference of Gaussian)算子探测图像中的感兴趣区域,然后用尺度不变特征描述器(Scale Invariant Feature Transform, SIFT)[11]描述这些感兴趣区域,最后对这些区域的SIFT特征进行聚类来生成视觉词汇。本文中视觉词汇表的生成也参照该方法。文献[6,8]中的研究表明,在场景分类中基于网格采样方法的性能要优于基于特征点探测或者感兴趣区域探测的方法,因此,本文采用网格稠密采样来获得感兴趣区域。具体来说,视觉词汇表的生成过程如下:

1)对训练图像进行网格稠密采样,从而得到相应的网格采样点,本文使用的网格采样间隔为8像素;

2)对每一个网格采样点提取其周围区域的SIFT特征来描述该网格采样点,在本文中使用网格采样点周围16×16的区域来计算SIFT特征,SIFT特征用该区域梯度方向直方图表示,是一个128维向量;

3)使用K-均值聚类算法对训练图像集上的所有网格采样点的SIFT特征进行聚类,每个聚类中心对应一个视觉词汇,从而生成一个由 K 个视觉词汇构成的视觉词汇表。

2.2 构建图像分块的BOW描述

图像中有意义区域对于场景分类有着重要的作用。人们获得图像区域的方式有很多,如图像分块、图像分割等。考虑到当前图像自动分割技术仍然无法得到令人满意的结果,因此本文采用对图像进行均匀分块的方式来生成图像的区域。本文采用 $N \times N$ 的固定网格来对图像进行均匀分块,这样每幅图像可以得到 $N \times N$ 个分块区域。对于每一个图像分块,我们分别构建其BOW描述。具体过程如下:

1)根据前面生成的视觉词汇表,将图像中每个网格采样点的SIFT特征对应到视觉词汇表中的一个视觉词汇。具体来说,就是用网格采样点的SIFT特征与视觉词汇表中的每个视觉词汇所对应SIFT特征进行比较,从而找出最相近的视觉词汇来表示网格采样点。

2)分别统计在每个分块区域中视觉词汇出现的频率,从而构建每个图像分块区域的BOW描述。

2.3 基于PLSA方法的分块潜在语义提取

在构建了图像区域BOW描述后,可以利用PLSA方法来进行分块潜在语义的发现。

将图像中的每个分块区域看作一个单独的文档,用 d 来表示,而视觉词汇就看作文档中的词汇,用 w 来表示,图像分块区域的潜在语义用 z 来表示,$n(d_i, w_j)$ 表示视觉词汇 w_j 在分块区域 d_j 中出现的频率。

基于PLSA方法的分块潜在语义提取可以分成两个阶段:

1)训练阶段:对由训练图像生成的所有图像分块集合,应用PLSA来进行训练,通过EM算法迭代直到收敛,从而得到 $P(w \mid z)$。这里 $P(w \mid z)$ 实际上就是分块潜在语义模型,它描述了在图像分块中潜在语义出现时视觉词汇的分布规律。

2)推断阶段:对测试图像的所有分块区域,保持 $P(w \mid z)$ 不变,同样用EM算法迭代直至收敛,从而得到每个分块区域的 $P(z \mid d)$。$P(z \mid d)$ 表示了分块区域具有潜在语义 z 的概率。

假设定义分块潜在语义的个数为 T,对于每一个分块区域 d_i,我们可以得到一个 T 维的特征向量 $[p(z_1 \mid d_i), \cdots, p(z_T \mid d_i)]$。考虑到分块潜在语义在空间上的分布也有助于图像场景分类,因此,我们最终将图像 $N \times N$ 个分块的 T 维特征向量连接为一个 $N \times N \times T$ 维向量 $[p(z_1 \mid d_1), \cdots, p(z_T \mid d_1), \cdots, p(z_1 \mid d_{N \times N}), \cdots, p(z_T \mid d_{N \times N})]$,这就是我们定义的图像分块潜在语义特征。在获得图像分块潜在语义特征后,可以通过构建SVM分类器模型来对图像进行场景分类。

3 实验结果与分析

本文实验中使用的图像数据是由文献[6]提供的13类场景图像数据集。该数据集共包括13类自然场景(括号中的数字代表每一类的序号):卧室(1)、海岸(2)、森林(3)、高速公路(4)、城市(5)、厨房(6)、起居室(7)、山脉(8)、办公室(9)、乡村(10)、街道(11)、郊区住房(12)、高楼(13)。每一类场景由200 ~ 400幅图像组成,图像的平均大小为300×250,并且所有图像都是灰度图像。这些图像的来源包

括 Corel 图像集、个人相册和 Google 图像搜索引擎。

由于要对图像进行均匀分块来生成图像分块,而不同大小图像所得到的分块大小也不相同,这可能会对最终的分块潜在语义模型生成带来一定影响。因此,在本文的实验中,将图像的大小都归一为 256 × 256 像素来进行处理。

整个图像数据库被随机分为训练集和测试集两部分。首先从每类场景的图像中随机选择 100 幅图像加入训练集,然后将每类场景中剩余的图像加入测试集。视觉词汇表一旦训练好是固定的,我们通过从训练图像集合中随机选择小部分图像来构建视觉词汇表。综合考虑实现的效率,本文中使用视觉词汇数目为 600,而区域潜在语义数目为 40,后面的实验都是采用相同的设置。

由于涉及到多类场景的分类问题,本文中的多类分类器是通过 one-vs-rest 的方式来构造:对每一类场景都学习得到一个区分它和其他场景类别的 SVM 分类器,本文中采用径向基函数(Radial Basis Function,RBF)作为 SVM 分类器的核函数;然后对每一个测试图像,分别计算每一个场景分类器对其的输出置信值,置信值最高的就作为该图像所对应的场景类别。实验中对整个图像库进行了 10 次随机划分来生成相应的训练集和测试集图像,然后分别计算每次划分的分类准确率。10 次划分得到的分类准确率的均值将作为最终的平均分类准确率。

文中首先比较了使用不同分块方式对于场景分类结果的影响。采用不同分块方式所得的平均分类准确率如表 1 所示。不难看出,采用不同分块方式对于分类性能有一定影响。由于我们已经对图像大小进行归一化,因此,采用不同分块方式得到的图像分块大小也不一样。从表 1 中可以发现,当分块大小过小(8 × 8)或者过大(1 × 1)时,分类的性能较差。其原因在于分块过小时,分块中往往没有包含有意义的语义内容,而分块过大则包含了太多的语义内容,从而使得分块潜在语义模型不够准确。后面的实验中都将采用 4 × 4 的分块来划分图像。

表 1　不同分块方式的分类性能比较

分块大小	平均分类准确性/%
1 × 1	72.9
2 × 2	75.6
4 × 4	78.1
8 × 8	74.3

在此基础上还比较了本文方法与文献[6 - 8]中方法在 13 类场景图像上的平均分类准确率,如表 2 所示。不难看出,通过引入分块区域潜在语义和其空间分布,本文方法相比于其他三类方法具有更高的准确率。图 1 给出了本文方法得到的混淆矩阵。混淆矩阵的 X 和 Y 轴分别表示场景类别,第 i 行 j 列的值表示第 i 类图像被分类为第 j 类图像的比例,混淆矩阵对角线上元素的值代表了每类场景的分类准确率。从图 1 中不难看出,在 13 类场景中,卧室(1)、厨房(6)、起居室(7)是最容易错分的类别。因为这三类场景都是室内的场景,其局部特征空间分布变化比较大,没有一定的规律,用本文中的分块潜在语义特征比较难于分类,需要结合其他图像特征来进行场景分类。

文中的视觉词汇表和分块潜在语义模型的生成是离线建立的。在 P4 2.0 GHz 和 1 GB 内存的机器上,花费了大约12 h 的时间来生成。对一幅测试图像进行分类,本文方法需要 20 s左右的时间,而文献[8]方法需要 15 s 左右的时间。本文方法因为需要对图像所有的分块应用 PLSA 来拟合得到分块的潜在语义分布,同时在特征描述中加入了空间分布信息,因此分类时需要花费更多的计算时间。

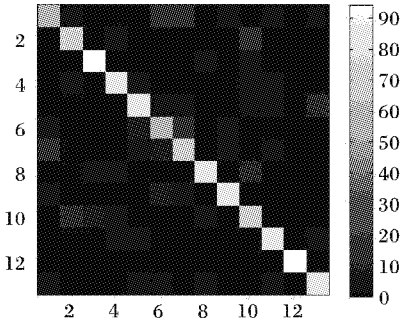


图 1　13 类场景分类的混淆矩阵

表 2　本文方法与其他方法的比较

方法	平均分类准确性/%
文献[6]的方法	65.2
文献[7]的方法	66.5
文献[8]的方法	73.4
本文方法	78.1

4　结语

本文提出了一种基于无监督方式获取图像分块潜在语义进行场景分类的方法。该方法首先用 PLSA 从未标注的图像分块区域集合中发现分块潜在语义模型,然后根据分块潜在语义模型找出图像中各分块区域潜在语义的出现情况,从而构建图像的分块潜在语义特征来进行场景分类。与文献[5]中方法相比,本文方法不需要进行区域语义的手工标注,而是以无监督的方式直接从数据中发现图像分块的潜在语义模型。与文献[6 - 8]中方法相比,本文方法既考虑了区域语义信息,也考虑了区域语义的空间分布特性,因而能更好地对场景建模。在 13 类场景上的实验也表明,该方法具有更好的性能。

然而,当前研究仍然存在一些不足,对一些类别的场景用本文特征还不能取得很好的分类效果。因此,如何提取出更有效的特征来描述场景图像,如何综合多种特征来进行场景分类都是下一步值得研究的方向。同时,本文中的分块潜在语义只在一个尺度层次上进行了计算,而且分块大小对于分类性能有一定影响,因此,多尺度分块潜在语义表示也是一个值得研究的内容。

参考文献:

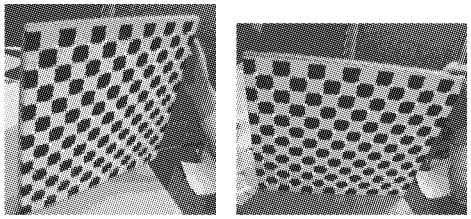
[1] VAILAYA A, FIGUEIREDO A, JAIN A, *et al*. Image classification for content-based indexing [J]. IEEE Transactions on Image Processing, 2001, 10(1): 117 - 129.

[2] CHANG E, GOH K, SYCHAY G, *et al*. CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(1): 26 - 38.

[3] SZUMMER M, PICARD R W. Indoor-outdoor image classification [C]// IEEE International Workshop on Content-based Access of Image and Video Databases in conjunction with ICCV'98. Washington: IEEE Computer Society, 1998: 42 - 50.

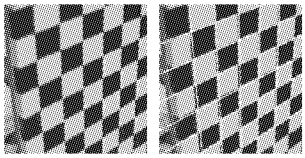
(下转第 1542 页)

点提取的平均误差比较如表1所示,图6为部分检验图样。可见本文的算法与传统的Harris算法具有较为接近的检测准确度。而方法2的检测结果误差很大,难以接受,究其原因,是镜头径向失真导致的,如图7所示。



(a) (b)

图6 用于准确性检验的部分图样

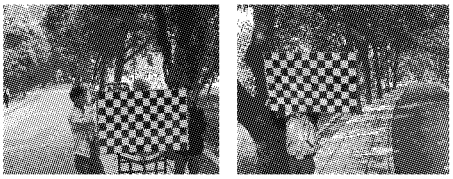


(a) (b)

图7 径向失真造成的检测误差

3.2 鲁棒性验证

在实验室以外,光线及图样的背景均不可控,单纯的经典角点检测算子与基于Hough变换的检测算法均显得无能为力。图8所示为户外场景下获取的检验图样,表2所示为本文算法与其他方法对于户外图样的方格点检测的效果的比较。可见,本文算法对于干扰较多的户外图样具有很高的鲁棒性。方法1在能够检出棋盘图样上所有的方格点以外,在树丛、道路等处多检出一百多个干扰点。至于方法2,由于图样中线条比较杂乱,直线的交点均不能正确落在真实的棋盘方格点上,因此检出的54点均为错误的点,而真实的154个点均没有被检测出来。



(a) (b)

图8 干扰较多的户外图样

表2 对户外图样的方格点检测效果比较(共154个点)

	错检点数	漏检点数
方法1	167	0
方法2	54	154
本文方法	0	0

4 结语

在分析现有方格角点检测算法的优缺点的基础上,结合棋盘标定物的几何特征,提出了一种从粗到精的方格角点检测及坐标精确提取的算法。先粗略确定方格角点所在区域,再在区域内根据对称性特征实现方格角点坐标的精确提取。实验表明,算法能鲁棒地对方格角点进行自动检测和精确提取。

参考文献:

[1] CLARKE T A, FRYER J G. The development of camera calibration methods and models [J]. Photogrammetric Record, 1998, 16(91): 51 - 66.

[2] 马颂德,李毅. 计算机视觉中摄像机定标综述[J]. 自动化学报, 2000, 26(1): 43 - 55.

[3] 谭晓军,沈伟,郭志豪. 机器人立体视觉模块的故障诊断[J]. 华中科技大学学报: 自然科学版, 2005, 33(6): 86 - 88.

[4] TSAI R Y. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses [J]. IEEE Journal of Robotics and Automation, 1987, 3(4): 323 - 344.

[5] ZHANG Z. A flexible new technique for camera calibration [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330 - 1334.

[6] HEIKKILÄ J, SILVÉN O. A four-step camera calibration procedure with implicit image correction [C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97). San Juan, Puerto Rico: [s. n.], 1997, 1: 1106 - 1112.

[7] 谭晓军,余志,李军. 一种改进的立体摄像机标定方法[J]. 测绘学报, 2006, 35(2): 138 - 142.

[8] Camera calibration toolbox for Matlab [EB/OL]. [2007 - 02 - 08]. http://www.vision.caltech.edu/bouguetj/calib_doc/.

[9] 胡海峰, 熊银根. 一种基于两次Radon变换检测棋盘方格点的新算法[J]. 中山大学学报, 2003, 42(2): 23 - 26.

[10] BEAUDET P R. Rotationally invariant image operators [C]// Proceedings of the 4th International Joint Conference on Pattern Recognition. Tokyo: IAPR, 1978: 579 - 583.

[11] KITCHEN L, ROSENFELD A. Gray-level corner detection [J]. Pattern Recognition Letters, 1982, 1(2): 95 - 102.

[12] HARRIS C G. Determination of ego-motion from matched points [C]// Proceedings of Alvey Vision Conference. Cambridge: [s. n.], 1987: 189 - 192.

[13] HARRIS C G, STEPHENS M. A combined corner and edge detector [C]// Proceedings of the 4th Alvey Vision Conference. Manchester: [s. n.], 1988: 147 - 152.

[14] BRACEWELL R N. Two-dimensional imaging [M]. Englewood Cliffs, NJ: Prentice Hall, 1995: 505 - 537.

(上接第1539页)

[4] 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.

[5] VOGEL J, SCHIELE B. Natural scene retrieval based on a semantic modeling step [C]// International Conference on Image and Video Retrieval, LNCS 3115. Berlin: Springer-Verlag, 2004: 207 - 215.

[6] LI FEI-FEI, PERONA P. A Bayesian hierarchical model for learning natural scene categories [C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2005, 524 - 531.

[7] QUELHAS P, MONAY F, ODOBEZ J-M, *et al*. Modeling scenes with local descriptors and latent aspects [C]// Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05). Washington: IEEE Computer Society, 2005: 883 - 890.

[8] BOSCH A, ZISSERMAN A, MUNOZ X. Scene classification via pLSA [C]// European Conference on Computer Vision, LNCS 3954. Berlin: Springer-Verlag, 2006: 517 - 530.

[9] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning, 2001, 41(2): 177 - 196.

[10] BLEI D, NG A, JORDAN M. Latent dirichlet allocation [J/OL]. Journal of Machine Learning Research, 2003, 3(7): 993 - 1022.

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