用于图像场景分类的空间视觉词袋模型

王宇新[1] 郭 禾[2] 何昌钦[1] 冯 振[1,2] 贾 棋[2]
(大连理工大学计算机科学与技术学院 大连 116023)[1] (大连理工大学软件学院 大连 116620)[2]

摘 要 以传统的词袋模型为基础,根据同类场景图像具有空间相似性的特点,提出了一种用于图像场景分类的空间视觉词袋模型。首先将图像进行不同等级的空间划分,针对对应空间子区域进行特征提取和k均值聚类,形成该区域的视觉关键词,进而构建整个训练图像集的空间视觉词典。进行场景识别时,将所有空间子区域的视觉关键词连接成一个全局特征向量进行相似度计算。最终的场景分类结果使用V1滤波器和PACT两种特征在支持向量机LIBSVM上获得。

关键词 场景分类,词袋,空间聚类,空间视觉词典,支持向量机
中图法分类号 TP301 **文献标识码** A

Bag of Spatial Visual Words Model for Scene Classification

WANG Yu-xin[1] GUO He[2] HE Chang-qin[1] FENG Zhen[1,2] JIA Qi[2]
(School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China)[1]
(School of Software, Dalian University of Technology, Dalian 116620, China)[2]

Abstract An approach to recognize scene categories by means of a novel model named bag of spatial visual words was proposed. Images were hierarchically divided into sub regions and the spatial visual vocabulary was constructed by grouping the low-level features collected from every corresponding spatial sub region into a specified number of clusters using k-means algorithm. To recognize the category of a scene, the visual vocabulary distributions of all spatial sub regions were concatenated to form a global feature vector. The classification result was obtained using LIBSVM and two kinds of features were used in the experiments: "V1-like" filters and PACT features.

Keywords Scene classification, Bag of words, Spatial clustering, Spatial visual vocabulary, SVM

1 引言

随着数码设备的普及和信息存储与传输技术的快速发展,图像数据发生爆炸性增长。如何用计算机对大量且不断增加的图像进行分析和理解,成为一项越来越紧迫的任务。因此基于内容的检索技术已成为国内外研究的热点,并成为21世纪初必须攻克的关键技术之一[1]。

本文将重点关注图像场景识别和分类问题。传统的场景分类方法通常使用色彩、纹理和形状等图像底层视觉特征直接与监督学习方法相结合进行图像场景分类[2];或者对场景中的目标进行有效的分析,以完成场景的整体识别,具有代表性的如王涛、胡事民和孙家广院士提出的基于颜色-空间特征的图像检索方法[3];或者采用文本主题模型的方法将图像分类到不同的语义类别中:将图像的局部不变特征聚类为一组视觉词汇,并用词袋(Bag of words)模型来表示,最后用LDA(Latent Dirichlet Allocation)[4]或PLSA(Probabilistic Latent Smantic Analysis)[5]等主题分析模型找出图像的潜在语义和最可能属于的主题,从而完成场景分类。

人类视觉感知的一个显著特点是能够很快掌握一幅复杂图像所表达的含义。Mary Potter通过实验证明,仅仅观察一组快速的图像流,观察者也能识别出每一幅图像的语义类别和一些图像中的对象及其属性[6]。这种通过快速(大约200ms)观察图像所获得的视觉和语义信息称为图像的gist[7]。在拍摄照片时,摄影师总是尽可能把能反映图像gist或者语义的对象和特征显示在图像中心。这一拍摄习惯使得大多数针对同类目标的图像都有相同的拍摄角度,即这些图像具有空间相似度。例如,许多城市的图像是这样一种景观:高楼下面连接着人行横道,而顶上是蓝天;高速公路是一个很大的平面向水平线方向延伸,期间充满了一些凹凸。这样,如果我们把图像划分为一些空间子区域,对应的子区域内就应该有相似的特征,如图1所示。

上述主题分析模型是根据图像中视觉词汇出现的总体情况进行分类的,既没有考虑视觉词汇在空间的分布特点,也没有利用图像中区域语义构成的上下文信息,而这些是决不能被忽视的。空间金字塔[8]模型的提出更是给了研究者非常大的启示。本文提出了一种用于图像场景识别的空间视觉词袋

到稿日期:2010-07-08 返修日期:2010-10-09
王宇新(1973—),男,博士生,讲师,CCF会员,主要研究方向为图像处理、计算机系统结构,E-mail:wyx@dlut.edu.cn;郭 禾(1955—),男,教授,博士生导师,CCF高级会员,主要研究方向为计算机系统结构、计算机视觉;何昌钦(1986—),男,硕士生,主要研究方向为图像识别;冯 振(1987—),男,博士生,主要研究方向为基于内容的图像检索;贾 棋(1983—),女,博士生,讲师,主要研究方向为计算机视觉、人工神经网络。

模型。

以传统的词袋模型为基础，引入图像空间信息，在将图像进行不同等级的空间划分后，针对同一空间子区域进行特征提取并聚类，以形成该区域的视觉关键字，进而构建整个训练图像集的空间视觉词典。进行场景识别时，把所有空间子区域的视觉关键词连接起来形成一个全局特征向量进行相似度计算，以获得最终的场景分类结果。

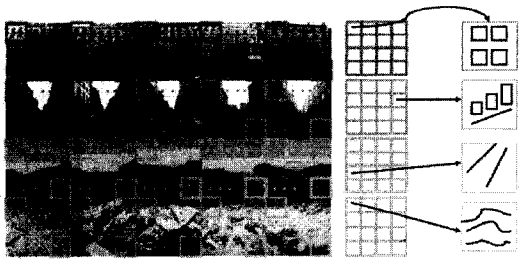

图 1　空间子区域具有相似性示例

2　空间视觉词袋模型

2.1　词袋模型与视觉词典

词袋(Bag of words)模型原是自然语言处理领域用于文本信息检索和文本分类的技术[9]。用它做图像表示模型(我们称之为视觉词袋模型)，需要将二维的图像信息映射成视觉关键词集合，这样既保存了图像的局部特征又有效地压缩了图像的描述[10]。

为了使用视觉词袋模型，首先要在学习阶段建立视觉关键词的集合：在训练图像集中，提取出所有局部特征，然后对这些特征进行聚类，得到的结果是训练集中的普遍特征。我们称这些原型特征为“视觉词典”。

2.2　空间聚类

视觉词袋模型在从训练图像集的图片中提取低级特征后，需要使用一种无监督的算法，如 K 均值算法[11]，对这些低级别特征进行给定聚类中心数目的聚类。给定一组观察值的序列$(x_1, x_2, \cdots, x_n)$，这里，每一个观察值都是一个 d 维的实值向量。K 均值聚类的目标是划分这 n 个观察值到 k 个序列里 $S=\{S_1, S_2, \cdots, S_k\}$ $(k<n)$，见式(1)，其中 μ_i 是 S_i 的均值。

$$\operatorname{argmin} \sum_{i=1}^{k} \sum_{x_j \in S_i} \| x_j - \mu_i \| \tag{1}$$

通过把一个聚类中心当作一个视觉关键词，就能把每一个从图像中提取的特征映射到它最接近的视觉词典上，并且能把图像表示为一个视觉词典上的直方图特征。

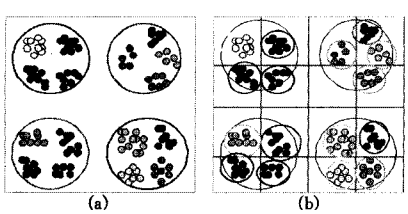

图 2　不同范围的聚类示例

在视觉词袋模型中，聚类是最重要的。词典的语义准确性对最后的识别率有着直接的影响，它取决于词典中的特征是否是图像中最普遍的特征。如果聚类中心数目一定，那么聚类的范围越广，语义信息越容易丢失。例如，在图 2 中，(a)是直接聚类而没有考虑聚类的范围，而(b)是划分聚类范围为空间的子区域。从(a)和(b)的区别可以很直观地看出，对应子区域内的原型特征能通过在空间子区域内聚类得到。

2.3　空间视觉词典

我们已经知道大多数图像具有空间相似度，并且在空间子区域内聚类能得到属于对应空间子区域的原型特征。因此，提出一种“空间视觉词袋”模型，它是视觉词袋模型的扩展。具体来说，首先有层次地把图像进行空间划分(如图 3 所示)，再把空间对应子区域聚集在一起，构建属于对应空间子区域的空间词典，过程如图 4 所示。

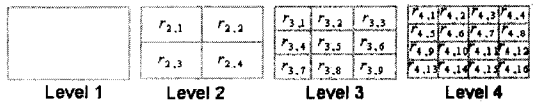

图 3　不同层次的图像空间划分

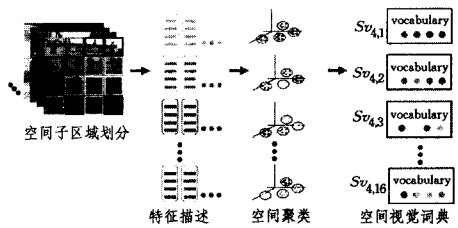

图 4　空间视觉词典的构建

形式上，定义 $r_{l,i}$ 为第 i 个空间子区域的 l 级划分，Sn_l 为 l 级划分下空间子区域的数目，这样就得到 $Sn_l = l^2$ $(l=1, \cdots, 4)$。同时可以定义 $Sv_{l,i}$ 为 l 级划分下第 i 个空间子区域内的视觉词典，$i=1, \cdots, l^2$。

2.4　基于空间视觉词袋模型的场景分类过程

一旦得到了每一个空间子区域的视觉词典，视觉词袋表示就能通过常规的方式得到。对于每一幅图像，空间子区域内的每一个像素的特征表示被投影到相应空间词典的一个单词通道上。每个子区域的第 k 直方图通过统计有多少像素在单词通道 k 上得到，这个直方图表示在空间子区域内的视觉词典的词频，见式(2)。

$$H(r_{l,i})(k) = \sum_{j \in r_{l,i}} I[T(j)=k] \tag{2}$$

式中，$I[\cdots]$是指示函数，$T(j)$是返回映射到像素 j 的关键词[12]。通过这种方式，就引入了空间子区域的空间相似信息。

为了识别一个场景的类别，需要把所有空间子区域的视觉关键词连接起来，得到一个全局特征向量。最终的分类识别率通过支持向量机来获得，整个过程如图 5 所示。

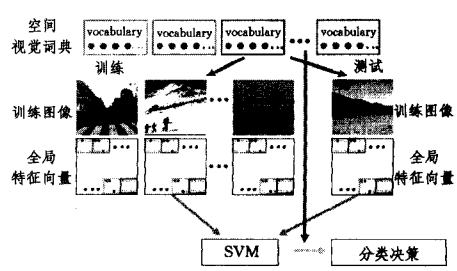

图 5　空间视觉词袋模型分类过程

2.5　图像间的相似度度量

当两个对应的子区域已经通过相同的对应空间词典计算出其视觉词袋时，我们通过比较其特征之间的距离来衡量其

相似度。

令 $H(r_{l,i}^{I_1})$ 和 $H(r_{l,i}^{I_2})$ 分别为两幅不同图像 I_1 和 I_2 在 l 级别划分下的第 i 个子区域的视觉词袋表示。我们用 χ^2 相似度来衡量二者之间的距离，计算方法见式(3)，相似度示例如图 6 所示。

$$\chi^2(H(r_{l,i}^{I_1}),H(r_{l,i}^{I_2}))=\frac{1}{2}\sum_{k=1}^{|Sv_{l,i}|}\frac{[H(r_{l,i}^{I_1})(k)-H(r_{l,i}^{I_2})(k)]^2}{H(r_{l,i}^{I_1})(k)+H(r_{l,i}^{I_2})(k)} \tag{3}$$

在 l 级别划分下，图像 I_1 和 I_2 之间的距离用式(4)进行计算。

$$D_l(I_1,I_2)=\sum_{i=1}^{Sv_l}\chi^2(H(r_{l,i}^{I_1}),H(r_{l,i}^{I_2})) \tag{4}$$

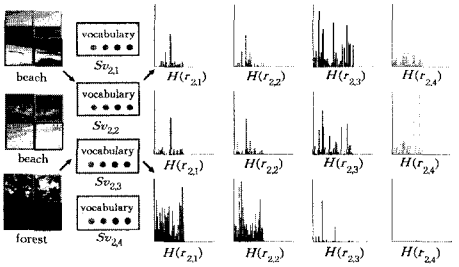

图 6　不同图像间的相似度示例

3　特征提取

本文提出的是一种能适用于各种类型特征的通用框架，本节将简单介绍两种用在实验章节的图像特征：表述人类的纹理识别力的“V1-like”滤波器和中心变换的主成分分析 PACT。

3.1　V1 滤波器

V1 区域是灵长类动物的大脑皮层的视觉区域，是最简单、最早的视觉皮层区。它在处理静态和动态对象信息以及在模式识别中具有重要作用。许多多尺度滤波器模型能描述人类的纹理识别力，这些滤波器满足了 V1 区域中简单皮层细胞感受野的定义[13]。

本文模型是基于空间对应子区域的相似度的，我们是从人类认知识别角度找到这一特性的。所以首先选择 V1-like 滤波器作为一种特征类型，并采用高斯函数的一阶导数和二阶导数去构造积分对[14]，见式(5)。

$$\begin{aligned} f_{\text{odd}}(x,y)&=G'_{\sigma_1}(y)G_{\sigma_2}(x)\\ f_{\text{even}}(x,y)&=G''_{\sigma_1}(y)G_{\sigma_2}(x)\end{aligned} \tag{5}$$

式中，$G_i(x)$ 表示一个具有标准差为 σ 的高斯函数。$\sigma_2:\sigma_1$ 是衡量滤波器延伸率的一个标准。滤波器组具有 3 个尺度自由度、6 个角度自由度，它们比 gist 滤波器具有更广的感受野。

3.2　PACT

为了更好地阐述模型的优点，同时采用另外一种完全不同于 V1 滤波器的特征 PACT(Principal Component Analysis on CT histograms)，即中心变换直方图的主成分分析[15]。主成分分析 PCA (Principal component analysis) 是一种统计分析方法，它能从多元事物中提取主要因素，从而反映事物的本质。

PACT 中的中心变换比较了中心像素与周围 8 像素的强度值，举例如下：

30	50	70	1	1	0
60	50	20	=>0		1=>$(11001010)_2$=>CT=202
80	30	70	0	1	0

通过强度值比较得到的 8 位数可以以任何顺序组合起来(采用从上到下、从左到右的顺序)，得到一个 8 位二进制数，与它对应的十进制数范围在[0,255]区间。如图 7(a)、(b)所示，变换后的图像不仅包含了全局特征，而且捕捉到了局部细节特征。

(a)原始图像　　(b)变换后图像

图 7　PACT 变换示例

视觉词袋模型采用聚类来获得词典，词典中的单词就是图像集中最普遍的特征。而我们的模型是基于空间相似度的，换句话说，就是想找出图像集的空间对应子区域内的最普遍的特征。PACT 中 PCA 提取了中心变换直方图分布中最重要的特征，从一定角度上说，和视觉词典具有类似的含义，所以我们想获得空间 PCA 的实验结果来加以比较。

4　实验与分析

4.1　实验 1

首先使用 Oliva 和 Torralba[16] 提供的图像数据库做实验测试，此数据库包含 8 类场景图片，每类中有 200 到 400 幅、大小为 256 * 256 像素的图像，如图 8 所示。分类识别率通过支持向量机 LIBSVM[17] 来获得。

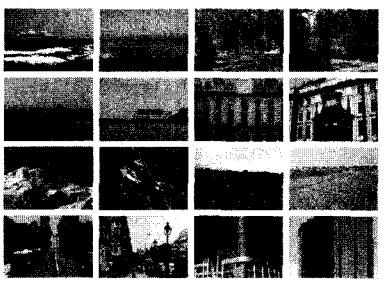

图 8　Oliva 和 Torralba 图像数据库

表 1 显示了使用 V1 滤波器作为基本提取特征，每类场景用 100 幅图像做训练，其余图像做测试时的实验结果，L (Level)表示空间划分级别，Sv 表示每个子区域内聚类中心数量(视觉关键词数)，做对比的是经典的词袋模型方法。

表 1　V1 特征的实验 1 结果

	V1-like filters(|Sv|=100)		V1-like filters(|Sv|=200)	
L	经典词袋	空间视觉词袋	经典词袋	空间视觉词袋
1	58.1568		59.4456	
2		60.5932		61.5466
3		61.8644		60.6992
4		63.4004		60.4873

表 1 中可以看出无论空间怎样划分，聚类中心有多少，本文方法都比传统的不含空间信息的词袋方法的识别率要高。当子区域内聚类中心数合理时(如 100)，空间划分越细，场景

识别率越高。尤其当 $L=4$ 时，空间相似度被最大程度地发掘出来；而当聚类中心数偏多时（如 200），空间划分程度影响就不大了。因此，我们认为空间划分带来的较粗粒度的几何因素比更多的视觉关键词具有更高的辨别力。

空间词典中的视觉关键词能够捕捉子区域内普遍特征的广义上的词汇，而 PCA 能够从多元事物中提取主要因素从而反映事物的本质。我们同样在空间视觉词典框架内计算每个子区域的 PCA 特征。表 2 显示了使用 PACT 的实验结果，可以看到空间 PCA 方法的识别率远高于经典 PACT 方法。

表 2　PACT 特征的实验 1 结果

	PACT(PCA=40)	
L	经典算法	空间 PCA
1	66.6843	
2		73.9230
3		74.1702
4		76.4301

4.2　实验 2

Caltech-101[18] 是由 Li Fei-Fei 等构建的包含 101 个物品分类（如人脸、飞机、古物、钢琴等）共 9146 幅图像的图像库，如图 9 所示，很多类型的物品图片也具有空间相似性。

图 9　Caltech-101 图像数据库

按照 Svetlana Lazebnik[8] 的方法在 Caltech-101 图像库上做物体识别的实验：每个类别中训练 30 幅图片，测试图片是每类 50 幅。表 3 给出空间划分级别 $L=4$ 时使用 V1-like 和 PACT 特征时的平均识别率，可以看出空间相似度被发掘出来后识别率大大提高。

表 3　实验 2 结果

	V1-like filters(|Sv|=100)		PACT(PCA=40)	
L	经典词袋	空间视觉词典	经典算法	空间 PCA
1	18.7565	—	13.9206	—
4	—	30.7081	—	21.0017

结束语　本文基于传统的词袋模型提出了一种空间视觉词袋模型。图像被划分子区域，并计算每个子区域的视觉关键词以挖掘图像间的空间相似度，最终构建训练图像库的空间视觉词典。基于空间视觉词袋模型的图像场景分类通过支持向量机 LIBSVM 在两种不同图像库上进行了实际应用，相比经典的算法，本文方法较明显地提升了识别率。

虽然目前的图像空间划分取得了较好的结果，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。

参 考 文 献

[1] 李向阳，庄越挺，潘云鹤. 基于内容的图像检索技术与系统[J]. 计算机研究与发展，2001，38(3)：344-354

[2] 高隽，谢昭. 图像理解理论与方法[M]. 北京：科学出版社，2009：399

[3] 王涛，胡事民，孙家广. 基于颜色-空间特征的图像检索[J]. 软件学报，2002，13(10)：2031-2036

[4] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022

[5] Bosch A, Zisserman A, Munoz X. Scene classification via plsa [C]// Proceedings of the European Conference on Computer Vision. Graz, Austria, 2006: 517-530

[6] Potter M C. Short-term conceptual memory for pictures [J]. Journal of Experimental Psychology: Human Learning and Memory, 1976, 2(5): 509-522

[7] Oliva A. Gist of the Scene[J]. The Encyclopedia of Neurobiology of Attention, 2005: 251-256

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

[9] Lewis D. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval[C]// Proceedings of 10th European Conference on Machine Learning (ECML-98). Chemnitz, DE: Springer Verlag, Heidelberg, DE, 1998: 4-15

[10] 李远宁，刘汀，蒋树强，等. 基于"bag of words"的视频匹配方法[J]. 通信学报，2007，28(12)：147-151

[11] Hartigan J, Wang M. A k-means clustering algorithm [J]. Applied Statistics, 1979, 28: 100-108

[12] Malik J, Belongie S, Leung T, et al. Contour and texture analysis for image segmentation[J]. International Journal of Computer Vision, 2001, 43(1): 7-27

[13] DeValois R L, DeValois K K. Spatial vision[M]. Oxford: Oxford University Press, 1990

[14] Renninger L W, Malik J. When is scene identification just texture recognition? [J]. Vision Research, 2004, 44(19): 2301-2311

[15] Wu Jianixn, Rehg J M. Where am I: Place instance and category recognition using spatial PACT[C]// IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA, 2008: 1-8

[16] 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: 23-36

[17] Chang C, Chang C, Lin C J. LIBSVM: a library for support vector machines[CP/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001

[18] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories[C]// IEEE CVPR 2004, Workshop on Generative-Model Based Vision. 2004